



**Calhoun: The NPS Institutional Archive**  
**DSpace Repository**

---

Theses and Dissertations

1. Thesis and Dissertation Collection, all items

---

2012-06

# Utilizing Twitter to Locate or Track an Object of Interest

Nauta, Jeremy T.

Monterey, California. Naval Postgraduate School

---

<http://hdl.handle.net/10945/7391>

---

*Downloaded from NPS Archive: Calhoun*



Calhoun is the Naval Postgraduate School's public access digital repository for research materials and institutional publications created by the NPS community. Calhoun is named for Professor of Mathematics Guy K. Calhoun, NPS's first appointed -- and published -- scholarly author.

**Dudley Knox Library / Naval Postgraduate School**  
**411 Dyer Road / 1 University Circle**  
**Monterey, California USA 93943**

<http://www.nps.edu/library>



# **NAVAL POSTGRADUATE SCHOOL**

**MONTEREY, CALIFORNIA**

## **THESIS**

**UTILIZING TWITTER TO LOCATE OR TRACK AN  
OBJECT OF INTEREST**

by

Jeremy T. Nauta

June 2012

Thesis Advisor:  
Second Reader:

Gary Langford  
Thomas Huynh

**Approved for public release; distribution is unlimited**

THIS PAGE INTENTIONALLY LEFT BLANK

<b>REPORT DOCUMENTATION PAGE</b>			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.				
<b>1. AGENCY USE ONLY (Leave blank)</b>		<b>2. REPORT DATE</b> June 2012	<b>3. REPORT TYPE AND DATES COVERED</b> Master's Thesis	
<b>4. TITLE AND SUBTITLE</b> Utilizing Twitter to Locate or Track an Object of Interest			<b>5. FUNDING NUMBERS</b>	
<b>6. AUTHOR(S)</b> Jeremy T. Nauta				
<b>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</b> Naval Postgraduate School Monterey, CA 93943-5000			<b>8. PERFORMING ORGANIZATION REPORT NUMBER</b>	
<b>9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)</b> N/A			<b>10. SPONSORING/MONITORING AGENCY REPORT NUMBER</b>	
<b>11. SUPPLEMENTARY NOTES</b> The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number ____ N/A ____.				
<b>12a. DISTRIBUTION / AVAILABILITY STATEMENT</b> Approved for public release; distribution is unlimited			<b>12b. DISTRIBUTION CODE</b>	
<b>13. ABSTRACT (maximum 200 words)</b>  Data in online social networks can be used as a resource to locate persons of interest. The two key issues are the accuracy and the length of time to carry out the necessary categorization, correlation, and sifting. Literally millions of data items—most unintentionally prepared to facilitate analysis—are posted and made available through public data feeds. The lack of appropriate tools and schemas inhibit efficient identification and extraction of information. The broad applicability of locating persons of interest extends to humanitarian assistance and disaster-relief efforts, finding missing person(s), reconstructing movements of people, and prognosticating future movement of people. This research defines a method that was shown to be effective in utilizing social network data (Twitter) to locate and track a person of interest. A combination of C# programming language and structured query sequences was integrated with SQL to correlate and sort hundreds of thousands of data items.				
<b>14. SUBJECT TERMS</b> correlation, data mining, data manipulation, SQL, Twitter, tweet			<b>15. NUMBER OF PAGES</b> 113	
			<b>16. PRICE CODE</b>	
<b>17. SECURITY CLASSIFICATION OF REPORT</b> Unclassified	<b>18. SECURITY CLASSIFICATION OF THIS PAGE</b> Unclassified	<b>19. SECURITY CLASSIFICATION OF ABSTRACT</b> Unclassified	<b>20. LIMITATION OF ABSTRACT</b> UU	

NSN 7540-01-280-5500

Standard Form 298 (Rev. 8-98)  
Prescribed by ANSI Std. Z39.18

THIS PAGE INTENTIONALLY LEFT BLANK

**Approved for public release; distribution is unlimited**

**UTILIZING TWITTER TO LOCATE OR TRACK AN OBJECT OF INTEREST**

Jeremy T. Nauta  
Lieutenant, United States Navy  
B.S., San Diego State University, 2005

Submitted in partial fulfillment of the  
requirements for the degree of

**MASTER OF SCIENCE IN SYSTEMS ENGINEERING**

from the

**NAVAL POSTGRADUATE SCHOOL  
June 2012**

Author: Jeremy T. Nauta

Approved by: Gary Langford  
Thesis Advisor

Thomas Huynh  
Second Reader

Clifford Whitcomb  
Chair, Department of Systems Engineering

THIS PAGE INTENTIONALLY LEFT BLANK

## **ABSTRACT**

Data in online social networks can be used as a resource to locate persons of interest. The two key issues are the accuracy and the length of time to carry out the necessary categorization, correlation, and sifting. Literally millions of data items—most unintentionally prepared to facilitate analysis—are posted and made available through public data feeds. The lack of appropriate tools and schemas inhibit efficient identification and extraction of information. The broad applicability of locating persons of interest extends to humanitarian assistance and disaster-relief efforts, finding missing person(s), reconstructing movements of people, and prognosticating future movement of people. This research defines a method that was shown to be effective in utilizing social network data (Twitter) to locate and track a person of interest. A combination of C# programming language and structured query sequences was integrated with SQL to correlate and sort hundreds of thousands of data items.



THIS PAGE INTENTIONALLY LEFT BLANK

## TABLE OF CONTENTS

I.	INTRODUCTION.....	1
A.	OBJECTIVE .....	1
B.	HUMANITARIAN ASSISTANCE AND DISASTER-RELIEF OPERATION .....	4
C.	SOCIETY AS A SYSTEM .....	5
D.	ADVANTAGE OF THE STUDY .....	10
	1. Case Study .....	10
	2. Example .....	11
	3. Scenario .....	12
II.	STRUCTURE .....	15
A.	STRUCTURE OF THIS THESIS .....	15
B.	SCOPE .....	15
C.	BOUNDARIES .....	16
D.	LIMITATIONS .....	17
E.	ASSUMPTIONS .....	17
	1. Related Tweets.....	17
	2. Relevant Tweets.....	18
	3. Retweets .....	18
	4. Access to Source Code.....	19
	5. Continuity .....	19
	6. Society as a System .....	20
III.	BACKGROUND.....	21
A.	SOCIAL NETWORKS .....	21
B.	TWITTER .....	22
	1. Twitter Privacy .....	25
	2. Readers and Users .....	25
	3. Twitter Handles .....	26
	4. Tweet or Retweet .....	26
	5. Following.....	28
	6. Twitter APIs .....	28
	7. A Tweet .....	29
C.	GEOLOCATION.....	30
D.	XML PARSER .....	31
E.	DATABASE .....	32
F.	STRUCTURE QUERY LANGUAGE .....	33
G.	KEY WORD GENERATION.....	35
H.	FUNCTIONAL DECOMPOSITION.....	35
I.	CORRELATION .....	36
IV.	RELATED WORK.....	39

V.	METHOD.....	41
A.	RECORDING TWITTER TO A DATABASE TABLE .....	41
1.	Creating the Database .....	41
2.	Coding the Recorder .....	43
a.	Using Twitter Source Code .....	44
b.	Using the Twitter APIs .....	45
3.	Key Word Generation .....	46
B.	PROCESSING THE DATA .....	46
C.	CORRELATION PROCESS.....	47
D.	MONITORING .....	48
VI.	ANALYSIS .....	51
A.	INITIAL ANALYSIS.....	51
B.	IDENTIFYING CHANGES.....	52
C.	CONCENTRATED ANALYSIS .....	52
D.	DETERMING LOCATION .....	54
VII.	SUMMARY AND CONCLUSIONS .....	57
A.	SUMMARY .....	57
B.	CONCLUSION .....	57
	APPENDIX A: A TWITTER CASE STUDY .....	61
	APPENDIX B: JOPLIN TORNADO CASE STUDY .....	71
	APPENDIX C: MLB TWEET .....	75
A.	MLB TWEET IN TWITTER FORMAT .....	75
B.	SOURCE CODE.....	75
C.	SOURCE CODE IN XML FORMAT .....	78
	APPENDIX D: EXCERPTS ON CASE STUDIES.....	79
A.	EXCERPTS ON CASE STUDY RESEARCH FROM LANGFORD 2012A AND LANGFORD 2012B .....	79
	APPENDIX E: TRANGULATION .....	83
	LIST OF REFERENCES.....	85
	INITIAL DISTRIBUTION LIST .....	91

## LIST OF FIGURES

Figure 1.	Human Control System with Feedback Loop .....	7
Figure 2.	Closed Loop Control System.....	7
Figure 3.	Twitter Process (a User's Perspective Developed from Twitter Basics).....	8
Figure 4.	Tweet by @Mets at 12:02 PM – 21 May 2012 via Web.....	23
Figure 5.	Tweet by @SarahPallinUSA at 9:45 PM – 14 Sep 2010 via Twitter for BlackBerry® .....	24
Figure 6.	Tweet by @StateDept at 1:31 PM – 21 May 2012 via Web.....	25
Figure 7.	Frequency Comparison of Three Twitter Profiles .....	52
Figure 8.	Identifying an Event Through Frequency Comparison.....	53
Figure 9.	A tweet posted by Twitter handle @MLB at 11:45 am on 30 May 2012 .....	75

THIS PAGE INTENTIONALLY LEFT BLANK

## LIST OF TABLES

Table 1.	SQL Wildcards .....	34
----------	---------------------	----

THIS PAGE INTENTIONALLY LEFT BLANK

## LIST OF ACRONYMS AND ABBREVIATIONS

@	Method to call out usernames in Twitter and assign users an identity <sup>1</sup>
#	Hashtag; method of assigning an object an identity in Twitter <sup>1</sup>
C#	Representation for the computer program language C-sharp
DBMS	Database Management System
DCL	Data Control Language
DDL	Data Definition Language
DML	Data Manipulation Language
GPS	Global Positioning System
GUID	Globally Unique Identifier
HA/DR	Humanitarian Assistance/ Disaster Relief
HTTP	Hypertext Transfer Protocol
Reader	Non-registered Twitter user
Reply	A direct response to a social post
Retweet	A repost or a response to a social post on Twitter
RSS	Real Simple Syndicate
SQL	Structure Query Language
TOS	Terms of Service
TPS	tweets per second
tweet	Social post on Twitter <sup>1</sup>
tweeter	Registered Twitter user <sup>1</sup>
Twitter	An information, social network <sup>1</sup>
Username	The name selected by the user when registering with Twitter; becomes the Twitter handle when the preceded by @ sign <sup>1</sup>
XML	Extensible Markup Language

---

1. See <https://support.twitter.com/articles/166337-the-twitter-glossary>.



THIS PAGE INTENTIONALLY LEFT BLANK

## EXECUTIVE SUMMARY

Social networks can be used as a resource in various applications. The problem is that without an appropriate tool, extracting various types of information is problematic. This research introduces a method that defines and provides a tool to utilize a social network to monitor various activities, explicitly finding the location of or tracking an object of interest. Exploiting the aspects of the social network site Twitter, a process of monitoring and recording was conducted by formulating and building a schema based on the combination of C-sharp (C#) coded algorithms for categorization, correlation, and data sifting, and then enacted through structured query language (SQL). Using SQL, a method to quickly sort and correlate the information from Twitter was applied. Follow-on analysis of the information extracted showed that Twitter can be used to locate or track an object of interest, given certain conditions—for example, tweets contain a geolocation or correlated tweets contain a location that can be determined—that are inherent in the data sources and analysis techniques.

The hypothesis that the frequency of tweets and retweets was a direct indication of the change in the activity or movement of an object of interest was shown to be true. By converting Twitter posts, i.e., “tweets,” into a Real Simple Syndicate (RSS) feed that provides an alternative formatting capability to monitor Twitter, the use of Twitter to locate persons of interest was demonstrated. In part, the ability to geolocate an object is due to the license agreement for use of Twitter data. That license agreement stipulates that users assign certain rights to Twitter, including Twitter’s access to location-based information, which is part of the users’ tweets. Twitter’s business model continues to evolve and now stipulates the right of Twitter to sell certain levels of access to the Twitter data, including location information. The data shared in Twitter (in text format) can be correlated with structures that assign meaning in terms of location and temporal aspects of location (movement, for example). The relations that exist between the data within tweets can be identified, parsed, and correlated. Analysis of the

data contributing to correlation over time revealed a natural frequency of Twitter usage that can be set as a baseline of use, i.e., a reference from which changes can be monitored. Monitoring for changes in the tweet frequency is indicative of the change in the social activity of identified person of interests. With a change in frequency identified, analysis of the data provides a sufficiency of information to determine the location or tracking in geospatial terms to a degree of precision and accuracy reflected in the contents of the totality of data found within the tweets. That is to say, a single tweet might or might not have enough information content, but the information content increases with increasing numbers of tweets.

Fundamentally, the combination of frequency and quality of tweets indicates that some social networks contain useful information that allows the capability for locating and tracking certain individuals that have an affiliation with the social network. That network affiliation can be direct, as a user of the social network, or indirect, through users of the network who report on certain individuals. Both direct and indirect sources of data add to the effectiveness of the algorithms that resulted from the research in support of this thesis.

This research developed a tool to collect and utilize the data shared on the social network Twitter. The aim of the tool was to monitor the activities of an object of interest; specifically to determine whether the location and movement of a person of interest could be first, discerned; second, correlated with a pattern of tweets (and retweets); and third, extracted and made useful to a degree of accuracy and precision in a short time period. Exploiting the aspects of Twitter involved developing a process of monitoring and recording tweets through the combination of the programming language C# and SQL. The focus was to script an algorithm that compared Twitter data with key terms to be correlated. The key terms were generated using potentially useful key word generators developed for search engines. Using SQL with a database, a method was developed that sorted and correlated the data and information from Twitter. Follow-on analysis of the information extracted showed that Twitter can be used to locate or track an object of interest on the basis of queries that posed various baselines from which

changes from baselines were detected and measured. The precision and accuracy of location and tracking was determined to be predicated on the content of the tweets, the time of the tweets in relation to the actual events, the frequency of tweets (and retweets), and the number of independent users who tweet.

THIS PAGE INTENTIONALLY LEFT BLANK

## **ACKNOWLEDGMENTS**

Many thanks to LT Jonathan Allmond and LT Jamie Mason for allowing me to generate and develop programs based on the C# code developed together in previous projects. Their guidance and assistance while developing my C# code was invaluable and I would not have been successful without them.

THIS PAGE INTENTIONALLY LEFT BLANK

# **I. INTRODUCTION**

## **A. OBJECTIVE**

Social networks can be used as a resource in various applications, such as news updates, job research [1], marketing research [2], and peer reviews [3]. The problem is that without the appropriate set of tools, extracting information from a social network is problematic in terms of validation or timing. Unless the information comes from a known, credible source, it must be validated. Likewise, the timing of the concatenation of information has to be investigated to determine age and provenance of the data. In particular, one attribute of the social network Twitter was analyzed. This research introduced a method that defined and provided a tool to utilize a social network to monitor various activities reported by the Twitter users. Explicitly, this thesis focused on finding the location of and tracking of a person of interest. The key assumption that underlined this research is that people are communicative about items of interest to themselves and to others. The hypothesis that the frequency of tweets and retweets are a direct indication of the activity of an object of interest built on that assumption and provided this thesis with a specific focus. The information that can be found in a social network such as Twitter can be monitored and recorded in a database (depending on the user agreement between Twitter users and Twitter and the policies of Twitter in place at the time of tweeting). A Twitterer (or “tweeter”) is a Twitter user defined as an account holder who reads and posts tweets [4]. A “tweet” is a message posted to Twitter, and a reply, or “retweet,” is the sharing or spreading of a tweet after it has been posted [4]. Tweeting is the act of uploading a tweet.

The objective was to use the data shared in Twitter as the source of data that was then correlated and used to determine the location, or contribute to the tracking, of a person of interest. “To correlate” was to determine the relationship that exists within the data that is tweeted. For this thesis, a person of interest was



defined as a person whose whereabouts is unknown, and his or her location is needed for verification of safety or questioning [5]. The suitability of the software algorithm that allows for high performance and efficient data management [6] is essential when collecting and analyzing thousands or even millions of tweets. “Suitability” means fit for use, meets the objectives, satisfies the requirements of the stakeholders, and falls with the limitations set by the key stakeholders (e.g., cost, schedule, and policy). The issue is the processing speed and throughput to locate someone or track someone within a reasonably short period of time. While that timeframe for processing and analysis may not be real-time or even near-real-time, there are many applications of locating and tracking an individual that have shorter time requirements than a retrospective or historical meaning.

The scope of this thesis was to investigate the technology issues relating to the processing and analysis of Twitter data, recognizing that the confounding factors inherent in the type of data and types of accesses to the data may be more significant than the actual algorithm that was designed to correlate data. Some of those confounding factors are security features established for the protection of privacy, continuous changes made to improve access, physical network limitations, and the potential commercial wealth of the data. Despite the confounding factors, the algorithm developed was employed to determine whether it is possible to locate an object of interest using Twitter data.

The following methods are used for this research and development, starting with the use of a computer program to convert tweets and retweets into a Real Simple Syndicate (RSS) feed. The format of the RSS feed provides the capability to record the data from Twitter into a database. The data is presented as dataset recorded to pre-established tables within the database. Recording the feed into a database allows the data to be placed into clusters, which are subsets of the data based on similarities within the dataset [7]. Each table represents a different cluster of data. A single database with several tables simplifies the structure of the database, while providing flexibility to expand with new tables, as new clusters are determined. Using the computer language Structure Query

Language (SQL), which is ideal for the manipulation or modification of databases [6], the data was correlated within and between the tables. Correlating data is the act of determining the relationship exists between data [8]. Correlating the data obtained from Twitter results in a correlation point in the datasets. Analyzing the data that contributes to the correlation point over time enables the possibility to determine a baseline frequency, i.e., the frequency around which the data is shared on Twitter, of tweets and retweets. A key step for this thesis process is to monitor the correlation point for changes in the frequency of tweets and retweets around the base frequency determined, which signifies the Twitter activity correlating to the object of interest is shifting. Analysis of the data associated with the change led to location information.

Several problems have to be solved before this method can be applied. In order to follow an object of interest in Twitter the object needs to be identified. Twitter objects are currently identified either with a Twitter handle (@Twitter) or with a hashtag (#Twitter). The Twitter handle is made up of the username preceded by the symbol @. A hashtag is an “organically” created keyword—a word that a user preceded with “#” to stand out [9]. Objects identified in Twitter are useful for monitoring people or items that have a recognized set of identifiers.

If the object is not identified in Twitter (that is, it does not have an assigned Twitter handle or hashtag) a method is required to identify the object. This thesis developed a method using search query to create an identity. The query term becomes the identity, which can be searched for, recorded and monitored as if it were an actual Twitter handle. To determine the best query to find an object of interest, the method of key word generation was employed. Typical key words were proper names, city names or abbreviations, nicknames, travel, and airport codes. Using SQL and query terms as the identity, the data in Twitter can be recorded and clustered.

Twitter allows and encourages its users to select their own usernames. By inspection, there was no guarantee there was a relationship between the username being monitored and the object of interest. Likewise, when a key word

is generated and queried, multiple results were displayed. A method to correlate records of interest includes a process to filter the unnecessary and unrelated records that result from the query.

The next process is to extract the tweets related to the object of interest from the noise in Twitter. The other information that is shared on the social network is considered noise for the purposes of this research. Noise is defined as data that does not result in a correlation between the data and the object's identity. For example, to locate John Doe, a query is executed. All of the information related to a John Doe is extracted and correlated. A portion of the information correlated may not pertain to the John Doe of interest, or the information correlated contains data that is irrelevant to the activity of John Doe. To eliminate the noise, the information was filtered. Removal of noise in this manner was facilitated by the use of the standard tools found in SQL.

## **B. HUMANITARIAN ASSISTANCE AND DISASTER-RELIEF OPERATION**

Humanitarian Assistance and Disaster Relief (HA/DR) operations require situational awareness and information. The Navy breaks the requirements for HA/DR operation into the following categories: 1) Information and Situational Awareness 2) Command, Control, Communications, Computers and Intelligence (C4I) 3) Logistics 4) Health Service Support, and 5) Personnel, Skills and Capabilities [10]. Though there is no specified time limit to obtain the information, it is a goal to obtain as much information regarding the situation as soon as possible [10]. The five categories can be viewed as the minimum factors of information required. Twitter provides a method to communicate and share information.<sup>2</sup> C4I includes the status of communications infrastructure and the sharing of information, which is currently being done with Twitter. The method of correlation developed in this thesis can relate capabilities with services required by location, or assist in establishing a logistical infrastructure, knowing the

---

2. See <https://twitter.com/about>.

location of a need and the location of the resources to fill the need. Furthermore, the organizers of the HA/DR operations require specific information about activities within the area of concern and the people that are affected, such as the number of injuries or the amount of damage. Analysis of correlated tweets has the potential to provide HA/DR organizers with situational awareness needed to formulate a respond. The information needs to be shared with organizers and decision-makers to provide definitive location and movement information on people, characterize the situation sufficiently to support movement of materiel, and assess the timing and phasing of the various needs.

While social networks can play a vital role in HA/DR operations because of the specific information contained within the contents of messages and postings from participating users, both individual needs and collective needs may be identifiable. Social networks can also be used to pass this information between organizers and decision makers, as well as to inform the users of the network. The users can respond with additional information or make updates to information already in the network. The model for this participatory action is termed “Peer Production” [11]. In the event of a disaster, people share information on social networks at will [12], [13]. It is up to the users, as readers, to interpret the information that is posted by the emergency response personnel. Currently, there are no methods in place to extract the data from the social networks for a definitive purpose, i.e., locating and linking capabilities with needs.

### **C. SOCIETY AS A SYSTEM**

A system can be built with sensors that monitor certain traits of the system. As the traits change, the control processes that are designed to provide stability to the system make adjustments to improve the local performance or the meta-performances of the system. For example, in package tracking systems used by companies to manage shipping logistics, sensors are placed in known locations. When a package is scanned by the sensor, the controls system updates the package’s status with that location. An operator can then call up the

package via its serial tracking number and the location of the package is displayed. Operators check the status of packages to ensure timely and accurate delivery. In general, these control processes are in place to sustain the system or protect the system. For example the controls may send an alert to the operator if a package is scanned at an improper location to minimize the fallout from a package being lost [14].

The purpose of the sensor-feedback-control processes is to automate the monitoring process of the system and to deliver the information to the system controls. Figure 1 is an example how the human brain acts as a sensor-feedback-control loop, and Figure 2 is an example of an automated sensor-feedback-control loop. The system controls use information passed from the sensor and determines an appropriate response to ensure the system continues to operate within the limits of the control processes. The controls and sensor of the system take the place of an operator, a person whose sole purpose is to watch the system and adjust the system as needed. As users of a social network share information about themselves by tweeting, users who retweet are adjusting the social network system. Applying control theory to society, a social network can be viewed as the sensor for the system. The implications of a social network being a system sensor for society is a fundamental for finding a person of interest. However, there are no control processes in the system (i.e., the social network system) to monitor changes in the system's transactions. Those transactions are the postings and readings of tweets. There is no specific property, trait, or attribute identified within the structure of processes of a social network by which to monitor the activities of these transactions. Figure 3 illustrates the Twitter process. After users access Twitter, they may choose to create an account with Twitter, log into an existing account if they have already registered, or not register and only participate as a reader without the capability of submitting feedback. Once a user registers and/or logs into an account, they may choose to tweet, retweet, or read tweets and retweets found on twitter. The

process continues until the user chooses to terminate the session. Figures 1 to 3 all show a looping process that can be monitored.

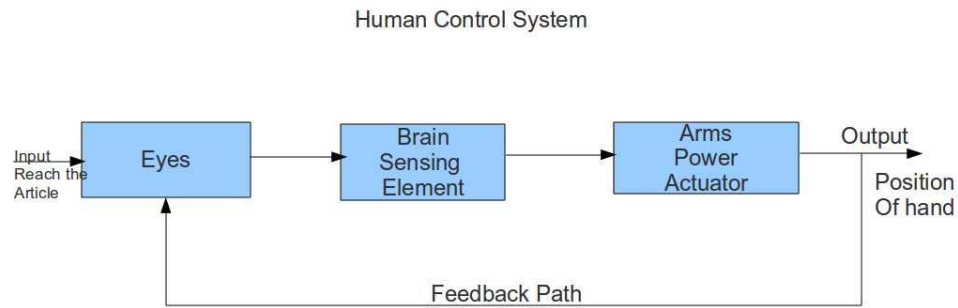


Figure 1. Human Control System with Feedback Loop<sup>3</sup>

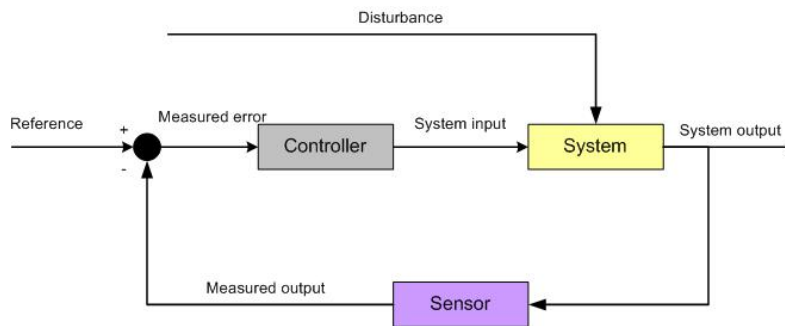


Figure 2. Closed Loop Control System<sup>4</sup>

3. See <http://instrumentationandcontrollers.blogspot.com/2010/11/feedback-principle.html>.

4. See <http://www.logitags.com/cibet/controltheory.html>.

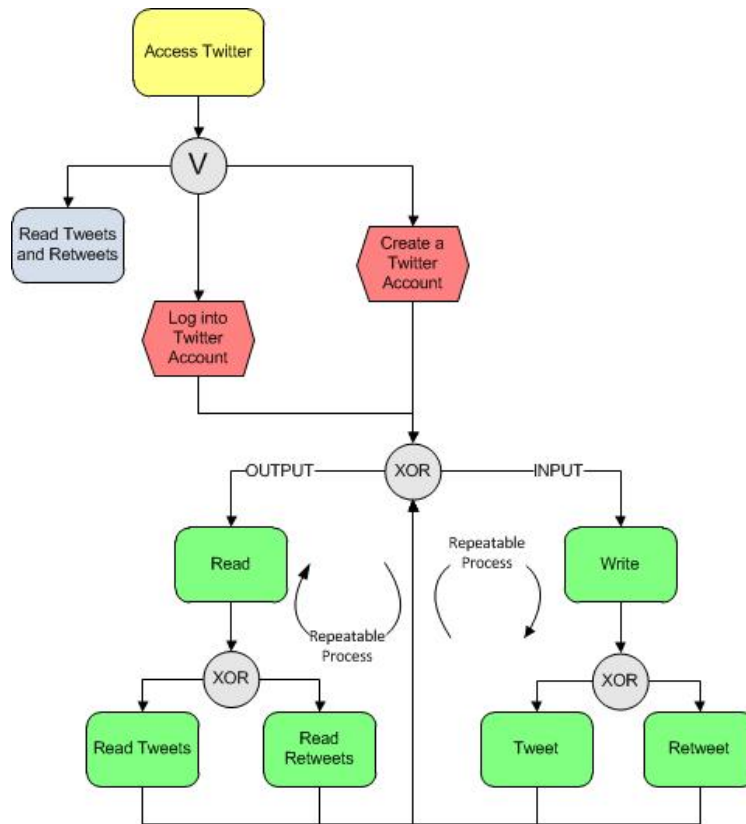


Figure 3. Twitter Process (a User's Perspective Developed from Twitter Basics<sup>5</sup>)

The user monitors the social network and inputs a response that in turn affects the social network. The influence is measurable based on how a user shares, recommends, or creates content on the network [15]. The current method of extracting information from social networks is through search for information conducted by a user (operator). In this method, the user analyzes the information obtained from the initial search then submits a second query (i.e., responds to the initial search results). This is a repetitive, manual process of accessing social network data using iterative or recursive approaches. The processing of data is both time-consuming and burdensome with regard to the amount of analytical knowledge required to sort through the responses to queries, analyze the data,

5. See <http://support.twitter.com/groups/31-twitter-basics>.

evaluate the data in terms of the task, and then organize the data into a meaningful structure for use in relating the results to the task.

Utilizing social networks in their current state for HA/DR operations is problematic. Organizers and decision makers need accurate and properly formatted information to satisfy the demands for quickly assessing the situation, making informed decisions, and then carrying through with action. Developing a process to monitor the social network and generate a response quickly by HA/DR organizers is beneficial because planners and responders can ill afford the time to find and process data required for their work.

An event can be identified by analyzing the user-inputted data and the way that data is changed or enhanced through the processes that are operative within the structures of the social network. These structures include the processes of forming and storing data, the processes employed by the users, and the processes whereby data can be extracted and evaluated according to a particular task. The persons involved can be considered objects of interest, which can be inferred using the information obtained [16]. Their location can be determined using the method developed and discussed in this thesis. As soon as the event is known (i.e., the event of location of a person of interest), a response can be set in motion, minimizing the time for adequate decision making. A more appropriate response, enabled by geographic information, reduces the uncertainties inherent in assessing a stressful situation, thereby improving HA/DR operations. For the task of finding a person of interest, the same timeline, accuracy, and formatting of information is necessary.

The process of finding a person of interest is typically carried out in five steps: 1) gather background information on the person of interest 2) determine the circumstances under which the person of interest might be involved 3) posit scenarios that place the person of interest within those circumstances 4) characterize the types and kinds of reporting through social media networks that



might be suggestive of the relations between circumstances and the scenarios and 5) correlate the data to discern patterns and references (expanding the list developed by [5]).

#### **D. ADVANTAGE OF THE STUDY**

Monitoring Twitter for an object of interest and identifying the activity of that object is a key step in using social networks as a useful sensor for the society. Twitter feeds can be used to monitor the health of a society or the changes in a society. However, before this control process can be developed the social network has to have useful information that can be analyzed. There needs to be a trait in the social network that can be identified so that monitoring it provides the necessary control. Correlation of the information pertaining to a trait provided the capability of analysis. The process of monitoring and analyzing Twitter can be applied to venues such as HA/DR operation, emergency response or like-kind issues.

##### **1. Case Study**

Case study methods (Appendix D: Excerpts on Case Studies) can provide a structured way to increase the understanding of how (in this case) a social artifact such social media can be transformed into a tool that is useful for assessing a situation that may require the following: some type of response by authorities, obtain practice in identifying strategic issues and patterns of behavior, improving judgment of decision makers, and gaining exposure to situations that otherwise might not be a part of one's experience [17]. The goal of case study research is to establish the parameters by which the data from one case study can be generalized to aid in understanding of like-kind situations [18].

There are two representative case studies that emphasize the importance of social networks and the need for an appropriate tool to integrate them. The first case study (Appendix A) was carried out in March 2010 by Kaitlin LaCasse and Laura Quinn, employees of Idealware. Idealware is a 501(c)(3) nonprofit that

helps nonprofits make software decisions by providing resources about software. In this case study, Idealware first summarized how a company can establish a Twitter account for free. Second, Idealware used the case study to explain how Twitter can be utilized to enrich the company. There are two important takeaways from the case study. The first item to note is the emphasis on the unique culture of Twitter, which revolves around sharing resources, but the difficulty involved without an appropriate tool. The second item of importance is the confirmation that Twitter can be an effective tool in communications.

The second case study (Appendix B) was initiated by Dave Bourne in May of 2011. As the Corporate Communications Manager for the Scarborough Hospital in Toronto, Canada, he raised the question of integrating social media into the hospital's strategy for crisis communication following the tornado that occurred on May 22, 2011, in Joplin, Missouri. Bourne focused on how the staff of St. John's Medical Center in Joplin used social media to direct and inform family members looking for those injured in the tornado event. Responses to Bourne's question complete the case study. To follow up with his case study, Bourne developed an index to measure the reputation of his hospital in Toronto. His presentation "Taking the pulse of healthcare social media: A prescription for measuring reputation" is important to look at because it provides an example of how the interaction of people in social networks influence the network and the users of the network.

## **2. Example**

Consider teen music star Justin Bieber as a person of interest and that his location is to be derived from Twitter. Adapting the steps explained in [19] and the process derived from [5], Twitter was used to find his location. High school student Cady Eimer established a website called [onelesslonelyprom.com](http://onelesslonelyprom.com) in January 2011 that hosted articles and personal videos in the hopes of having music star Bieber take her to senior prom. Though Bieber did not take her to prom, her hard work and dedication resulted in a date with Justin to the 2012

Billboard Music Awards in Las Vegas, where he won an award for Top Male Artist [20]. Eimer has amassed 21,000 tweets since establishing a Twitter account. She created a list in excess of 47,000 Twitter users whom she follows, and she is on the list of over 400 users that “follow” her [link to her page]. By reading her tweets, an understanding can be made for her fondness for the teen star by her expressions in her tweets regarding him.<sup>6</sup> She expresses in her tweets how much of a fan she is of the teen star. On May 19, 2012, her Twitter page exploded. Recording the tweets just from her page, it can be determined that she received a date request from Justin Bieber and accepted. Analysis of the tweets that followed show a pattern of tweets and corresponding events. These correlations of tweets related principally to date, billboard, award, Vegas, prom, and the list goes on. A thorough review of the tweets revealed she arrived in Las Vegas at 10:55 am on May 20, 2012, and left at 11:36 am on May 21, 2012. The tweets are arranged in chronological order, followed in sequence of her receiving the invitation to the Billboard Music Awards to her return to Virginia. By correlation, it was determined that Justin was also in Las Vegas during the period with a tweet, “rehearsal for the BILLBOARD AWARDS [sic] tonight.”<sup>7</sup> During the review and analysis of the information on Twitter, it was shown how tweets can be analyzed to find a person of interest.

### **3. Scenario**

The following account of an incident at a sporting event illustrates how social media can potentially be advantageous to the police. Los Angeles is a large metropolitan area in California with a population of 4 million people, covering 1,290 km<sup>2</sup>, with 10,000 police officers [21], and is the home of the Los Angeles Dodgers baseball team. In the case of Brian Stow, Dodger fans unleashed their anger by beating Stow, who was rooting for the opposing (and

---

6. See <https://twitter.com/#!/CadyEimer>.

7. Post available at <https://twitter.com/justinbieber/status/204261649450934272>.

winning) team after a Dodger's game in 2011. Emergency response personnel took nearly 15 minutes after the beating to reach him and provide medical attention [22].

Users of social networks are compelled to tweet and retweet. The social mechanism is that of sharing in near-real time any item that seems to the users to be interesting, noteworthy, or in some way relevant to their life. Users keep in touch with their “followers” through social networking as a means to eliminate the stored information that is moment-to-moment important, but might otherwise be overcome by subsequent events. The psychological reasons for wanting (and perhaps needing to tweet) are to satisfy the need to share and receive information, express an idea or talk about a new idea, and to simply smile. Extrapolating from social theory, humans are social animals, and sharing is a key aspect of that social fabric [23]. Within a few minutes of something that social network users want to share, they tweet. Within a few more minutes, the followers of that user retweet. Based on analysis of Twitter tweets and retweets in support of this thesis, monitoring a social network may have reduced the time for emergency response teams to arrive before the altercation began, or to treat Stow within a few minutes. Monitoring social network feeds allows skilled practitioners to infer possible changes in society on both the general level as an aggregation of behaviors, and specifically on an individual basis.

In this case, knowing that the Dodgers might lose the game to the rival SF Giants had angered fans, the police could have monitored social networks with the methods developed in this thesis to provide tools and processes to recognize the increased potential for violence, providing additional police presence in the area of angry fans.<sup>8</sup> As it is a recognized practice, providing an increased presence of officers in the area of the angry fans might have assisted in maintaining order and civility.

---

8. Historical Twitter records for this case could not be obtained, but other examples of tweets containing references to violence were found at <https://twitter.com/#!/search/saw%20fight>.

That the use of social media may improve our lives is suggested by this thesis. Increased police presence may have stopped the beating all together, or reduced the amount of time to respond to the fight that put Stow in the hospital. Knowing that changes in society are detectable and can be processed to elicit various actions might be helpful to those interested in improving society, e.g., marketing agencies and investors. It would allow those interested to foresee the outcome of their influence or help them make the right influence.

## **II. STRUCTURE**

### **A. STRUCTURE OF THIS THESIS**

This thesis presents a method to record and correlate the information obtained from the social network site Twitter. It identifies and monitors the frequency of tweets and retweets of the correlated information obtained. The changes found in the frequency of the information being posted were compared to confirmed sources of information to derive location and movement data from the evaluation of the changes. The process was applied using Twitter as a data source to confirm that Twitter can be monitored and used to provide information that is useful in the locating and tracking an object of interest.

This thesis begins with an explanation of a social network and why a social network is analogous to a system. The aim of the thesis to apply social data from Twitter given the role Twitter has as system sensor for society. The basics of Twitter and how it works are discussed. The next section of the thesis focuses on the analysis and correlation of data and information derived from tweets. tweets are tracked by the frequency of postings by users. Next, the thesis discusses how to recognize the changes in the frequency found and what those changes meant in reference to the task. The last section of the thesis focus on applying the methods and tool developed to confirm the usefulness of Twitter.

### **B. SCOPE**

The aim of the research was focused on “how” to develop a method for extracting and utilizing the required information from a social network. The scope covered the Twitter databases and access policies, the utilization processes juxtaposed against Twitter processes, and the utility of Twitter data for locating a person of interest. This thesis demonstrates the processes needed and analysis results of tweets and retweets to determine the location of a high-profile person of interest.

### **C. BOUNDARIES**

There must be clearly defined boundaries to the correlation process. Correlation deals with bounds that delimit what is to be considered and excluded from matching reference queries with data. If there are not clearly defined or established boundaries, the process of compiling data and checking for correlation may result in nonsensical correlation functions. In other words, the process of checking one tweet against every other tweet continues unabated in time and number of tweets; the result is lack of correlation. Rather, a time-limited, content-limited schema was developed and experimentally validated against temporal partitioned datasets. The end result is for the correlation process to be both continuous in its ability to receive and process tweets, but interruptible without loss of data integrity or congruence. Therefore, a boundary was established using the “follow lists” [4] in Twitter to determine what information tweeted is needed and what can be ignored (either temporarily or permanently). The boundaries established allowed for a continuous processing to occur, without infinite or long extended runs without conclusive results.

The goal for congruence is to make the process of correlation all-inclusive, so no causal linkage is missed and no concatenating data is misconstrued. However, to be all-inclusive, the computing power required would be of such magnitude as to project large mainframe computing along with its commensurate costs. Limiting the computing power to a desktop commercial-off-the-shelf computer is therefore a stated requirement. Defining the specific capability of the desktop computer means determining the number of items to be a few items represented by one task that can be correlated near-simultaneously. Proper definition of a few is more than two but not several. For the process to work with high-profile people, the computing resources were designed to record the specifics about those individuals, as is demonstrated in Section V. Only the information available on the social network was recorded, reflecting both the scope of work and the boundaries designated for correlation.

## **D. LIMITATIONS**

The intention of this thesis is to develop the process of extracting data/information from Twitter via RSS feed and demonstrate the ability to develop a method of correlation utilizing the data to determine specifics, e.g., location, of an object of interest. Though it was desired, this thesis did NOT intend to develop an automated process of recording, analyzing, and displaying all the useful properties of Twitter. A fully automated process of recording RSS feeds, data correlation, and analysis for location determination was determined to be out of scope. Therefore, the limitation imposed on the research was to produce a manual system of processes, and show the benefits for automation as a recommendation for follow-on research.

## **E. ASSUMPTIONS**

### **1. Related Tweets**

Twitter is an open standard format, designed for public access. The common referent is termed a social network site. Social networks allow anyone or thing, e.g., computers or automated systems, to read and tweet on the site [24]. Twitter (the company) does not restrict its users in their use of Twitter resources, except for failing to follow the user agreement. Additionally, Twitter is distributed widely around the world, allowing users from nearly any location to publish their thoughts. A key aspect of the behavior of Twitter users is their predilection to post frequently on persons of interest and, importantly, to follow these persons of interest with great detail. That detail often includes data and information about the users' special interests about celebrities, and other persons of interest whom they desire to be associated with or find some newsworthy comment to make publically. Because of the freedom inherent in the structure and processes of Twitter, not all tweets were related to one another. Twitter was not indented to provide a record of every tweet with defined relationships to other



tweets, but rather to record and present tweets to other users to read and analyze. Users are free to interpret tweets and determine the relationship between tweets and social interaction.

## **2. Relevant Tweets**

Two key assumptions that underlie this research are focused on people: people are communicative about items of interest to themselves and to others, and people will use Twitter as a tool to communicate. While it must be assumed that the tweets being recorded are relevant, it is reasonable to assure a minimal amount of noise associated with the correlation schema, not all tweets are related, not even retweets. In the process of recording tweets and retweets, the schema incorporates a rule to only record relevant tweets. Relevant tweets are determined by similarities (identical words, phrases or names) in the data tweets contain. It is impossible to make this rule 100% effective. Therefore, it must be assumed that the tweets and retweets that are recorded are relevant, and those that have minimal or no relevance were dealt with by a subsequent processing schema.

## **3. Retweets**

There are instances of tweets with follow-on retweets that either individually or when combined are related to the task and therefore the person of interest. Twitter has very few rules as to when and how people post comments. A user may never post an original tweet; instead, the user may only retweet. Again, a retweet is the reuse of an original tweet or in response to a tweet. Likewise, there is not rule that said a user has to tweet or retweet. The user may choose to read only the tweets and retweets posted by others. An established relationship can be seen on Twitter by viewing the follow/follower lists. However, it is nearly impossible to sort every follow/follower list manually for every user on Twitter. This leads to the assumption that established relationships on Twitter contain tweets with follow on retweet.

#### **4. Access to Source Code**

The location of the Extensible Markup Language (XML) tag that is part of every Twitter post can be extracted for all tweets and retweets. Personal privacy is a concern in the United States (as well as other countries), and Twitter provides the opportunity for its users to “hide” their tweets from the public eye. The programmers behind Twitter are capable of extracting those tweets, but that requires user permission or court order [25]. This thesis focused on the open, publically accessible data and information that is contained within Twitter. This thesis assumes that Twitter will not change its privacy policy, which allows any registered user to access all publicly shared tweets. This policy assumes Twitter will continue to support application-programming interfaces (APIs) which are effective for adding functionality to software programs. APIs allow third-party programs to be developed which utilize the backbones of a robust program built to take advantage of the Twitter processes without the requirement for third-party developers to build the whole program structure and detailed processes from scratch. Twitter has committed to supporting APIs and allowing third party applications to use Twitter, as long as they are within the guidelines and user agreements for Twitter. If Twitter stops supporting APIs and disabling access to third party applications, the methods developed in this thesis will need to be revised or integrated into a new structure consistent with Twitter policy and processes.

#### **5. Continuity**

It is safe to assume there will be a modicum of continuity in communications during a disaster. As was the case in the Great Japanese Earthquake (2012), the infrastructure at Fukushima and other coastal cities was devastated. The people of Japan, especially in the affected area, continued to communicate via Twitter [12]. This communication indicated resilience of will and technology utilization during a catastrophic event, and is plausibly extrapolated to

be a valid assumption—that continuity in communications will be maintained. In order to rely on Twitter, in layman’s terms, social networking processes has are expected to be available with a sufficiency of connectivity and access to facilitate communications in the manner and timeliness that the users think of as typical for normal usage. The commercial advantage for Twitter (and all social networking processes) is to remain as the means of “usual” communications.

## **6. Society as a System**

The case is made that society operates like a system. There are attributes of society that are contained in a social network that can be monitored, much like the model of a sensor that provide input data to a system. As the attributes of society change, they are reflected in the usage and processes of social networks [16].

### **III. BACKGROUND**

#### **A. SOCIAL NETWORKS**

The interest in social networks is steadily increasing because information that is relevant to people's life is being updated constantly through postings. Social network sites are easy to access, navigate, and use. The required tools and equipment that are required may already be in use for business or home interests, and consequently there may be no explicit monetary expenses or training necessary. Social networks allow users to submit information for others to see. While participating members of the social network have access, even non-users of the network can arrange for access, according to access policies promulgated from time to time. Users share information about each other and about various topics by submitting information they know to the social network. User submissions of information enable social networks to contain vast amounts of information; information that consists of personal data, links between users and others, interests, and locations where tweets are sent and received. The social network is a storage depot of information and data architected and intended for the users to share. The full potential of the information that is collected by the network is just beginning to find a host of applications for research [1], marketing [2], and reviews [3], among other uses. The full potential of the social networking resources are just beginning to be explored. A social network, as a resource, has the potential to identify what new products will find acceptance in the marketplace, influence people's attitudes and behavior (Klout.com), provide survey information about particular issues, provide communications needed for allocating limited resources in time of emergency response, and aid in locating and tracking persons of interest.

For nearly a decade, social media and social networks have prospered, increasing in the number of users and the geographic coverage. Reference [26] depicts how "big" Twitter is. In spite of their popularity, some think that social

networks are a waste of time. “Social networking now eats up twice as much of our online time as any other activity. According to new statistics from Nielsen, sites like Facebook and Twitter now account for 22.7% of time spent on the web; the next closest activity is online games, which make up 10.2%” [27]. The larger the number of users there are and the more time users spend on social networks, the more relevant their networking becomes. Increasing the network expands the size and amount of information in the database of social interactions.

Social networks are valuable; their value rests in the people that utilize and share the information. However, estimating the value of social media is a challenge [28]. The point of advertising is twofold: reach new customers to pique their interest and to consider and buy certain products. The value of social networks to other organizations can be seen in the amount of money spent in advertisement. It is estimated that by 2012, 88% of U.S. companies will be using the social networks for marketing purposes [29]. As advertisers spend a portion of their budget to reach potential and current users of social networking sites to increase their business, they are able to analyze the information that is contained in the network to help pinpoint and tailor ads to their target customers.

## **B. TWITTER**

Twitter was founded as a social network to provide users with a simple micro-blogging platform. A micro-blog allows users to share information about themselves or items they find interesting or relevant in another context. Twitter allows people to share their lives through the World Wide Web.<sup>9</sup> Users can post statements, thoughts, and links and share others post as long as it is kept within 140 characters.

Figure 4 shows a tweet in reference to the New York Mets baseball team. This is a good example of how much information a tweet can contain. The tweet

---

9. From <https://twitter.com/about>.

holds the starting roster by position for the team. In addition to the roster, the tweet includes the day of the game, their opponent they are playing and in whose stadium they are playing. It is also worthy to note that abbreviations are used for standard baseball abbreviations for positions.



Figure 4. Tweet by @Mets at 12:02 PM – 21 May 2012 via Web

Figure 5 is an example of a tweet that uses made-up abbreviations or word substitutions to meet the 140 character limit of a tweet. With no prior knowledge of the word usage, it would be difficult to decipher what is being said. In this tweet, the number 4 is substituted for the word “for,” “bc” in place of “because,” and “make’em” short for “make them.” This is a good indication of how tweeters can be creative to meet the 140 character limit yet share their information.



Figure 5. Tweet by @SarahPallinUSA at 9:45 PM – 14 Sep 2010 via Twitter for BlackBerry®

Figure 6 is an example of link usage within a tweet to share information. Hypertext Transfer Protocol (HTTP) links can be very long. For example, “http://elections.nytimes.com/2012/primaries/candidates/mitt-romney?8qa” would take up 120 of the available 140 characters. The links implemented in Figure 6 are short because Twitter developed a method to shorten the length of links so they can be included in tweets and expand the amount of information that can be shared.



Figure 6. Tweet by @StateDept at 1:31 PM – 21 May 2012 via Web

## 1. Twitter Privacy

Twitter allows users to see and explore the information shared by other users. Unless a user protects his information, any user may view the information shared without receiving permission from the user who posted the information. As a confounding factor, privacy concerns of users is important to Twitter, and Twitter takes great effort to protect the privacy of its users [31].

## 2. Readers and Users

In this research, people who use Twitter are classified into two groups: Readers and Users. The first group (Readers) are people who are not registered with Twitter, nor have they accepted the terms of service [30] of Twitter. By investigation, readers participate in communications external to Twitter, but as of yet those are not quantified (as such, they are out of scope for this thesis). Readers (hereinafter defined as “readers,” where readers are not “users”) are able to access all unprotected data in Twitter but do not add information to the Twitter database directly. Whether a user or not, there is no difference when reading the information that is available on the Twitter network, but readers have



no means to provide any type of feedback or comment on the information, unless they register. In other words, readers cannot submit information to be shared with others; they can only read the information that is being shared.

The second group (Users) is composed of the registered users. They have complied with the TOS [30] and submitted personal information to Twitter. Twitter provides an account and grants users permission to tweet. Registered users (hereinafter defined as “users”) have the full capabilities to read the information being shared in the social network, and submit information to the social network to be shared.

### **3. Twitter Handles**

After a user registers with Twitter, creates an account, and obtains a Twitter handle, they may participate as a “reader” or a “user” who may submit information to the network. In order to obtain a Twitter handle (username), a user must submit personal information and accept the policies [30] before Twitter provides them with an account. The personal information includes full name and e-mail address by which to confirm personal identity. The Twitter handle consists of the username preceded with the symbol @. In most cases, the user selects his own unique username. If the desired username is taken, Twitter provides similar alternatives. The username can be composed of any combination of letters, numbers or symbols. The username does not have to be associated with the user’s name or personal information, although it often is. “@\_\_” is an example of a Twitter handle made of symbols. Figure “@MittRomney” is an example of a Twitter handle represented by the real name of the user and presidential candidate, Mitt Romney, and “@ThisDopeKid” is an example of a Twitter handle that the user chose that has no association with the user’s name.

### **4. Tweet or Retweet**

Tweeting is the act of posting a tweet or retweet on Twitter. A tweet is a posted message on Twitter, where a retweet is the posting of a message in

response to or in reply to either a tweet or a retweet. A retweet can also be the share or reposting of either a tweet or retweet. In either case, tweets and retweets have a maximum of 140 characters through which the user can submit information to the network. With only minimal network routing and server delay as soon as the user submits the tweet, the information is made available for all to see. Only a user can reply to tweets or submit retweet.

A retweet is different from a tweet for two reasons. First, a tweet is original, while a retweet is a response or a reply to a tweet. A retweet always follows a tweet. Second, a tweet is information with all new data, but a retweet is information that consists of new and old data. In part, the new data might be merely the fact that a particular user has retweeted. The retweet contains data from the original tweet in addition to the new data. The retweet contains old data because, when it is submitted, the retweet contains a reference to the tweet from which it originated. A retweet is similar to tweets in every other aspect. The retweet with the attached data is a record that is found on the Twitter network [31].

Twitter encourages users to register and establish a username to use the full features of the network, but Twitter does not make it a requirement. The information Twitter contains is public information because Twitter is an open format for public participation and expression with the limited restriction and registered users must agree to [30]. A user can choose to keep his tweets private, or he may choose for the information he posts to the network to go public. Twitter provides each user the option to share the posted information or make it private. The user may choose not to have the information he posts to the network “go public.” If a user chooses to make the information he submits private, only registered users that receive permission from the originating user can view the information.

## **5. Following**

There are an estimated 140 million active users on Twitter who tally a massive 340 million tweets per day [32]. If only 1% of the registered users tweeted per hour, that is still 1.4 million records worldwide. It is unlikely a user can manually follow or monitor all of these records. Twitter allows the user to select only the records posted from those they wish to see through a feature called “Follow” within the limits<sup>10</sup> set by Twitter [4]. Twitter users can enter a username they are interested in, then select to follow that username. If a user chooses not to follow any username or hashtag, then they will only see the information they submit to the network and all the responses to that information. When a user chooses to follow a username or hashtag, then the user will see all the tweets and retweets related to that username or hashtag including all of the tweets and retweets of the information they have submitted.

## **6. Twitter APIs**

Twitter is considered an information network with a slogan of “the fastest, simplest way to stay close to everything you care about.”<sup>11</sup> Three things are required for this to be true. First, there has to be some measurement of time to state that it is fast, and fast access to information that an organizer of a HA/DR operation cares about is good. Second, the access to and the processing of information from Twitter must be easier than having a user hunt for the required information, process the information and share it with the HA/DR organizers. Lastly, for this to be true the information must consist of data that is relevant or useful. This means there must be some method to form context of the information that is found in Twitter.

---

10. See <https://support.twitter.com/articles/15364-about-twitter-limits-update-api-dm-and-following>.

11. See <https://twitter.com/about>.

Twitter is an open-source social network with several built-in application programming interfaces (APIs) that allow the use of third-party applications [33]. A third-party application is a program not written by the primary program that adds functionality [34]. Twitter places limitations on the APIs to protect its users, but makes it very clear that APIs can and are used [30]. There are multiple applications available that use Twitter APIs, but most applications are developed so users can have tweets delivered directly to them while avoiding the logon process. An application can be installed on computer or mobile device. Once the application is set up by a user, his Twitter account is linked to the application, and the user no longer has to log into Twitter to read or respond to tweets. Handling of the tweets and retweets is done through the third-party application. Using the Twitter APIs, an application was developed to record tweets and retweets.

The first step is to convert the tweets into an RSS feed by exploiting the API features designed into Twitter. Since the Twitter structure was not designed to be an RSS feed, the API is a simple way to format tweets into an RSS feed [33]. The RSS feed is then recorded to create a history of tweets and retweets.

## **7. A Tweet**

A tweet has five basic parts. Behind the scene, Twitter attaches key data to form the context of the information. The data are: the date and time the tweet was submitted, the Twitter handle and full name of the person submitting the tweet, the source or location where the tweet was submitted, and an assigned Globally Unique Identifier (GUID). Lastly is the text of the tweet, which contains the subject. The subject of the tweet can be a person, place, event, etc. The tweet with the attached data is a record found in the Twitter network.

In no specific order of preference, the first part is the username. The user is the source of the tweet, the person responsible for entering the data, which is the second part of the tweet. The data is the text of the post. It is simply a piece

of data that may have no context, but context can be formed by using the other parts of the tweet. The third part is the time stamp. Every post contains a date-time stamp of when it was submitted. With the time stamp and the data, the post becomes information, and context begins to form. The last part of data is the geo-source. The geo-source can be a geolocation (that geolocates the tweeting device) or it can be a tweet source (indicated by content). When tweeting from a mobile device with a GPS receiver turned on, the geolocation of phone is the geo-source. However, Twitter and third-party applications offer users the option to hide the geolocation. When the geolocation is hidden, the software or platform from which the tweet originated is listed in place of the geolocation. In other words, if the tweeting platform is a desktop in a cybercafé in Dayton, Ohio, then that geolocation shows what software was used from a permanent computer station instead of the geolocation. The geo-source can still aid in the development of the context in either case because it provided distinct geographical coordinates or produced a binary indication of the post being made from fixed source or mobile platform. In the instance where the geo-source is the geolocation of the tweet as well as that of the person of interest, the tweet provides coordinates near where the tweet originates.

### **C. GEOLOCATION**

Geolocation locates an object of interest at a point on a map. The location is normally determined through Global Positioning System (GPS), but can be determined through an IP address, Wi-Fi network location, or self-reporting [35]. Late in 2009, Twitter announced that its platform developers were working to add geolocation features that allow users to include latitude and longitude to any tweet. Twitter recognized the potential advantage as a way to read tweets of the accounts around your location, not just the tweets of the people you follow [36]. The advantage of geolocation tags on tweets is the ability to see the location from where the tweet was submitted, indicating the location of the user.

Twitter can be accessed through the website Twitter.com. The geolocation tag attached when a tweet is submitted by a registered user through the website is the location the users entered when establishing the Twitter account. Web-based applications such as Echofon or Janetter provide various additional features that are not found on Twitter. Most importantly however, the applications provide users with the ability to tweet, but the geolocation tag attached when a tweet or retweet is submitted from a third-party application varies. The geolocation tag is either the location of the computer used or the name of the application. Other applications such as Twittrific or UberSocial are designed for mobile devices. These application are like the web-based application when it comes to tweeting, but it adds the benefit of “tweeting on the go.” The geolocation tag attached to a tweet when these applications are used also varies. When tweeting from a mobile device that has GPS enabled, the latitude and longitude is attached.

#### **D. XML PARSER**

Parsing is a process of information extraction that selectively extracts data to populate a database [37], [38]. An XML parser is a program that analyzes XML-formatted text, identifying segments of data in XML tags [39], [40]. An XML parser using the exact XML tag, e.g., <title> text /title, identifies the data. The XML tag opens with “<title>” and closes with “</title>.” The data is the text within the open and close XML tag [41]. The data, the text inside the open and close XML tag, can then be written to a database. The parser looks for the exact XML tag. If the exact XML tag is not known, the parser skips to the next instance of an XML tag to be parsed.

Modern web browsers have an XML parser built in [40]. Web applications like Twitter publish their code in hypertext markup language (HTML) and XML format so that the web browser will display the information properly. It is considered a bad idea and difficult to attempt to parse HTML [42]. Twitter uses both HTML and XML.

Developing an application that uses the information on Twitter required a parser. This researcher used an XML parser for simplicity vice attempting to parse HTML. When building the application, the XML parser was needed in order to interpret the XML code that was used [39]. C# contains a built-in parser, but requires the programmer to identify what is to be done with the data found within the XML tags. The data found within the XML tags for a tweet was written to a database for future analysis.

## **E. DATABASE**

Databases are essential for fast, efficient data processing [6]. Twitter works off a series of databases that are protected [31]. Developing a database is required to manage the data recorded from Twitter. Specific permission is required to access the Twitter databases, but the access is not free. Twitter economized its site and can sell the database records to whoever pays [43], [44]. However, access to the information in Twitter is free through Twitter's web application. A database had to be established that can retain the information recorded from Twitter. Using C#, an application can be written to record the data from Twitter by retrieving the data through the use of Twitter's web application and writing it to a database.

In order to use a database, a database management system (DBMS) such as Microsoft SQL Server [45] is required. Where the database is the data itself, the DBMS is what hosts the database and allows the use of SQL [46]. The DBMS provides the capability for the data manipulation, data clustering, data extraction, data addition, and data processing for the determination of correlation. The databases can be static or dynamic. Static databases are used for data that is permanent such as city names or proper names. Dynamic databases are used when there are changes in data, new data and expansion is required. The process of retrieving data from Twitter and recording it to a database is dynamic. As the dynamic databases are changed and there is expansion in the databases,

the DBMS maintains the relation between the data as it is recorded. The relation is the set of columns and rows that represent the data as it is recorded [47].

## **F. STRUCTURE QUERY LANGUAGE**

SQL is the optimal method of sorting information obtained from Twitter. SQL is the most predominant language for creating, configuring and querying databases [6]. It is comprised of Data Manipulation Language (DML), Data Definition Language (DDL), and Data Control Language (DCL) [48]. DCL deals with administrative controls for the database, which is essential when performing processes with databases [48], but does not provide added capabilities to this research. DDL provides the ability to change the database structure, and DML provides the ability to modify, add or extract information in a database [48]. DDL and DML are essential components of SQL for this research. It does not provide any insight to the discussion of this research by referring to the individual parts of SQL. For this research, the three components will be referred to as SQL.

The predominant features offered by SQL that are applicable to this research are the ability to sort, compare, and write data to a database or alter databases. These features are key for filtering or clustering data and the determination of correlation. Other features such as the ability to join or increment tables are useful, but do not provide an added benefit to this research. The language enables data to be written into a pre-established database with tables, but also provides the ability to create new or alter tables [47], [48]. The benefit of using SQL is the ability to query the database in a variety of methods, such as string searches and searches, using wildcards and the predominant features of SQL [6]. As long as a logical process for searching the database can be determined, SQL will allow it.

SQL also allows the manipulation of databases; however, caution is needed because the actions to manipulate the database cannot be undone. There is no limit to the number of manipulations that can be done to the database, but the manipulation that can be done must be done within the



constructs of the SQL. Manipulations outside the SQL must be handled via common code such as C-sharp (C#). This is referred to as embedded programming [6], [45]. SQL allows common code to execute the SQL code. This is beneficial because the code written in C# can be constructed to execute the database manipulations and the record permutations needed to analyze the data.

SQL wildcards are substitutes to find data within a database [49]. An SQL wildcard is a special character or character set that is substituted into a query to eliminate the necessity of precision. Table 1 contains a list of SQL wildcards and a description of the substitution they provide.

Table 1. SQL Wildcards<sup>12</sup>

Wildcard	Description
%	A substitute for zero or more characters
_	A substitute for exactly one character
[charlist]	Any single character in charlist
[^charlist] or [!charlist]	Any single character not in charlist

A drawback of SQL is the limited functions that can be performed to alter the values of the data in the database. SQL has some functions to manipulate the records (e.g., sum and average) [45], but it does not provide a way to change the text or format. These functions are only needed in the event the data must be converted to a usable format. With that said, it is important to write the data to the database in the required format. In this research, the data format is unknown or changes, and it may become necessary to alter the format of the data. To achieve the required functionality, SQL can be embedded in common program languages like C#.

---

12. See [http://w3schools.com/sql/sql\\_wildcards.asp](http://w3schools.com/sql/sql_wildcards.asp).

## **G. KEY WORD GENERATION**

Search engines such as Google generate revenue by selling keywords [50]. When a user enters the query word into the search area, the engine will display the links that are related to the key words. Advertising agencies will pay to have the link to their sponsor organization or company displayed at the top of a search list [51]. Key words are generated by following search engine click logs [52].

Reference [52] described how key words are generated and developed a method to generate key words from queries. Using the method found in [52], the key words to query when trying to locate a person of interest can be generated. The generated key words are then used in SQL to query the recorded Twitter feeds. All records that contain key words are written to a new table for further analysis. Inserting the key words generated into the query using SQL.

## **H. FUNCTIONAL DECOMPOSITION**

The purpose of functional decomposition is to understand what it means to locate or track an object of interest. In this research a functional decomposition of travel, locate and track was performed to identify what is needed to perform those actions. To travel is to move from one location to another, like travelling from home to work. Travel can be broken down into two key components. They are to schedule and to prepare. Each of these can be broken down with an explanation. To schedule travel is to block off a certain amount of time for the travel to be conducted. This can be the block of time to travel to work or this can be the block of a week for a vacation. In either situation, it needs to be allocated and blocked off to allow for travel to happen. To prepare for travel (in a functional sense) is to make sure that the traveler has everything required to travel.

A method of analysis of the Twitter data has to be developed and determination of what the Twitter data reveals will have to be verified using known information. This thesis will look at the action of travel. For this study, celebrities were monitored for their movement throughout the country. Much of

their travel and movement is released to public via their official website. The website information contains information on when they will be in a certain state or location. The information from their official website will be confirmed through various news reports and compared to the results determined using Twitter.

## **I. CORRELATION**

Correlation implies a relationship between two or more data points or datasets. When correlating data, the primary concern is finding a relationship that exists within the dataset [53]. A social network post or tweet is a data point that contains information, and it has the potential to be correlated with other tweets. If two randomly chosen tweets are related, that means the two tweets have some probability of a relationship; the two tweets contain related information. Tweets A and B correlate, then they have an intersection probability not equal to zero. In statistical terms  $P(\text{data point A} \cap \text{data point B}) > 0$ . The higher the probability of correlation between tweets,  $P(\text{tweet A} \cap \text{tweet B})$ , the greater the relationship between the two tweets. A higher probability of correlation led to higher precision.

To achieve a high probability of correlation, three attributes of the data were used. This is based on the principle of triangulation, a proven mathematical determinant for locating objects in two- and three-dimensional spaces [Appendix E]. Further, qualitative social research relies on triangulation to provide a preponderance of method to validate approach and results [Appendix E]. An accurate fix cannot be taken unless you have three or more points because it is a three-dimensional world. To accurately correlate tweets the time, subject and location should be correlated. However, a correlation of a single attribute can still be useful. A tweet contains a time stamp and geo-location. The third attribute to correlate is found in the subject of the tweet.

Twitter contains millions of tweets. Every tweet will not correlate, but each tweet contains a piece of information regarding an event. Using just one piece of data contained in the information of the tweet, i.e., object, the correlation with other tweets may be useless. It would be like calling a travel agent to reserve a

hotel room without giving a destination to look for or a date range for which to reserve the room. Time is a valuable asset, and the call would be a waste of time for both the travel agent and the caller. Likewise, there is little use to the information in the tweet when only correlating one piece of the data.

When correlating two pieces of data—e.g., object and location—the information starts to become useful. Adding the second correlation point to a location adds usefulness. In the case of the travel agent, the agent can now look at what is available for the location provided. For tweets, it is useful because the correlation between location and an object of interest indicates a relationship between the object of interest and that location. To the travel agent, the information provided is still vague. The agent may inform the caller when the location has its next availability, but this brings in the last date point: time.

Adding the third piece of data, knowing when to travel, completes the reservation. The travel agent can reserve a location for a person at a given time. Likewise, the third piece of data completes the correlation. The travel agent is correlating the caller's desire for a location and time with the availability of the location. Without a correlation between the availability and the persons wanting to travel, there would be no need for a reservation. Similarly, travel cannot occur for a person who wants to travel to a warm destination during the cold winter months, but never looks for a correlation between the destination and availability at the destination during the winter months, then there could be no travel.

THIS PAGE INTENTIONALLY LEFT BLANK

## IV. RELATED WORK

Organizations have looked into the usefulness of social networks. In an internationally sponsored research, Virginia Tech and the University of Campinas (Brazil) studied the pattern of communication and the information communicated using social media pertaining to the campus shooting in the University of Texas, Austin, on September 28, 2010 [13]. Unlike the research above, this thesis investigated the frequency of activity, not the pattern of communication. The basis for this research was the information communicated to establish a point that can be correlated then made available to monitor the activity around that the correlation point.

A number of studies have focused on finding trends in queries of datasets [54]. These studies query several datasets and analyze the trends found in the results of the queries. Similar to this research, queries were conducted on datasets. This research is relevant because finding trends supports the argument of a base frequency of Twitter usage. Identifying the changes in trends indicates a change in Twitter usage.

After the earthquake in Japan in 2011, tweets were exchanged on Twitter. An analysis of tweets immediately after the earthquake revealed that people in a disaster area tend to communicate directly with each other. On the other hand, people outside the area prefer spread information regarding the disaster area through the use of retweets [12]. Although [12] focused on who communicated, it was important to show that the network is robust and the infrastructure can be retained in the wake of a disaster. Additionally, [12] showed that users continue to share information throughout various events.

Companies like Geoemblem Technologies focus on geolocation information attached to information on the Internet. Their primary goal is to gather and store the information and for profit help organizations capitalized on opportunities in a geographical area of interest [55]. Geosemble gathers geolocation information found on social websites as well. In May of 2012, Geosemble released

GeoXray™ 3.0, which couples recorded geolocation information with satellite imagery [55]. Unlike Geosemble, this research is not just focusing on geolocation information but looking at the relevance of all the data within a tweet.

Emergency response teams just recently have begun to post information on social networks. Local police and emergency organizers are using social networks to notify the public (through the social networks as well as their traditional distribution channels). Police stations understand that people in their community monitor social networks and found that posting notifications to them provided a fast and inexpensive method to notify a segment of the public, in that case, a subset of the public with specific interests. Social networks provide a way for users to follow their local emergency response teams and provide the emergency response team with the ability to post information. The information is at the “fingertips” of the followers [56], [57]. The type of information that can be shared by city police over Twitter is analyzed [58].

## **V. METHOD**

### **A. RECORDING TWITTER TO A DATABASE TABLE**

Twitter is the name given to the social network program, but the focus of this study is on the tweet. Recording tweets was accomplished in a multiple-step method. An SQL database was established to store the tweets from Twitter, including all the data that is attached to the tweet. The tweets were converted to a format that can be recorded using a Twitter API. An RSS feed format was chosen for its simple structure that can be parsed. The API converts each part of a tweet into a segmented line of code to be displayed in an RSS feed. Using C#, an XML parser application was written to parse the RSS feed, separating each segment of the tweet and recording the data into a table within the SQL database.

Using Twitter, a query was executed and the results were displayed in the form of its source code. Analysis of the source code revealed two key features. The first feature noted was the XML tags being used. Knowing the XML tags was essential for creating the database and the Twitter recorder. The second feature displayed was the amount and type information that could be extracted from Twitter. To record Twitter, a database and tables were used to store the data. Also, a program to retrieve the information and “write” it to the data tables was completed and tested.

#### **1. Creating the Database**

A database was created using the DBMS Microsoft SQL Server 2008R. There was no certainty to the number of required databases or tables to contain the data recorded from Twitter. Using a DBMS allowed for the expansion of additional databases and tables without having to configure how the relationships between the databases and tables. An SQL server was used because it allowed for multithreading of the database. Multithreading (the ability to run multiple tasks



on the database at the same time) was necessary for this research to record the data while processing the data for correlation. The server also provided the benefit of executing table queries in the server management studio or in C#. Microsoft provides a temporary license for SQL Server 2008R (free of charge for academic programs<sup>13</sup>).

Initially, a single table was created using the XML tags found during the analysis of the source code of Twitter. For each XML tag in the source code, there was a designated column to record the data. The initial design was to use a single table that could be used to execute queries and form a correlation of data. The correlated data was then written to new table. Using a SQL database was essential to keep all the data together in the same form. It was advantageous to use the XML tags from the source code to create the table from because the information from Twitter is in the same order as the source code. This was found to be beneficial for debugging.

The first attempt to record Twitter to the database was met with errors, only one of which was caused by the database. Each tweet or retweet is limited to 140 characters; Twitter protocols that check the tweets will not accept any tweet if the character limit is exceeded. However, when recording retweets, it was found that Twitter adds a prefix. The prefix was “RT” plus the username. The recorder was set to record a string of unknown length, but the table created was designed to only hold a string of 140 characters (140 characters string length was chosen because the maximum length of a tweet). The actual string length exceeded 140 characters because of the added length of the prefix. The error was found during debugging, and prevented the recorder from writing the data to the table.

To correct the error found during debugging and prevent future string-length errors, the table was reconfigured for larger string lengths. When Twitter adds the prefix to the retweet, the string length is no longer restricted to 140

---

13. Thank you.

characters. Likewise, Twitter did not use a fixed string length when the prefix was added so there was not a fixed length to configure the table to. Though many tweets and retweets were still less than the 140-character limit including the prefix it was determined, a much larger string length would prevent future errors. The table was adjusted to accept strings up to 250 characters.

The disadvantage of configuring the table to use 250 characters was the requirement for additional computer memory. When configuring a table to a certain number of characters, the computer allocates memory for the predetermined string length. If the string written to the table does not use the full string length, it still uses the same allotted memory as a 250-character space. In larger databases or projects where memory is limited, this overcapacity would imply inefficiencies in search times. For this research, it was not a factor of concern.

## **2. Coding the Recorder**

A systems engineering approach was used to design the recorder. In particular, the focus on integration of both design and architecture were carried through with a set of requirements and performance measures. Functionally, the recorder was required to be integrated with Twitter and the SQL server. C# was chosen because of author's familiarity with the code and the abundance of available help (see acknowledgements), though several codes would provide the needed features. C# integrates with both Twitter and the SQL server, thereby supporting interoperability and integration; C# contains an XML parser and can execute SQL commands. The next step in the development of the recorder was to access Twitter using an API and view the source code. Knowing the XML tags from the source code was essential when developing the recorder. Next, the method of extracting the data tagged with the XML tags was completed using an XML parser. The last step was the development and testing of a method to write the extracted data to the data tables.

**a. *Using Twitter Source Code***

A user account was established with Twitter, which allowed full access to the features of Twitter. Abiding by the user agreement, the user was able to view and copy any tweet or retweet. The user established a “follow list” within the limits set by Twitter. Any tweet from a user that was on either list was then posted to the main Twitter page of the account established. The source code of the main page was extracted to record the reported information.

The initial attempt to record Twitter using the source code from the main Twitter page was met with many errors and ultimately unsuccessful. It was unsuccessful because neither the main Twitter page nor the source code updated automatically. This required a manual process whereby the user updated the main page to see if new tweets were posted. When considering the requirement for near-real-time access to facilitate geolocation, that update was determined to be a critical performance issues. If a new tweet was posted, a refresh of the source code was required to capture the new tweet.

Lastly, another confounding factor, the source code could not be parsed because the format was not conducive [42]. Appendix C contains an example of the source code for a single tweet; the original format used is hypertext markup language (HTML) which is similar to XML. Twitter developers used an efficient, adaptable method in HTML to post tweets to a user’s main Twitter page. Because each user can select whom they follow, each user’s main page can be different. To post tweets to the user’s main page, the developers used a HTML call method, but the method used only called the tweets that were associated with the user’s follow and follower list. The call methods did not write the XML tags in the source code, which prevented the recorder from extracting the tweet data based on the XML tags.

### ***b. Using the Twitter APIs***

It was determined the best way to record any information from Twitter was to use an API. The Twitter API converts the tweets to an RSS feed in a uniform manner. Again, the XML tags had to be determined in order to extract the information. To view the XML tags, the Twitter API was executed in a web browser, then the source code was viewed. The source code view revealed the data and the XML tags pertaining to each piece of data. Every tweet used the identical XML tags for the individual pieces of data when it was converted into the RSS feed. Additionally, using an API mitigates the confounding factor of access to the data.

The parser created in C# used the XML tags found in the source code. The parser followed the sequence of the RSS feed to find the XML tag that contained the data that needed to be recorded. The exact XML tag had to be known; otherwise, the parser would not find the tag. If the parser did not find the tag, the data would be skipped.

The first parser created was developed to capture tweets based on the Twitter handle. The results of the first parser showed an insufficient amount of information was recorded. Only 20 of the most recent tweets were recorded when recording was based on the Twitter handle. Further analysis showed that the API used only provided the twenty most recent tweets.

Twitter offers several APIs. To record all the information regarding an object of interest, three APIs were used. Each API converted the tweets to an RSS feed, but each feed was in a slightly different format. Looking at the source code for the three feeds showed that each had used similar XML tags. Another difference was the tags used were in a different order. Use of the Twitter API did not allow the tags to be altered or rearranged. Since each API converted the information to a slightly different format with different tags, a parser was created to handle each API.

Assigning a parser to each API simplified the process of recording the tweets. A single parser was attempted to handle all three feeds, but it became very complicated when trying to debug the program. A single executable program was never successfully completed. It was advantageous to use a parsing program for each API. Using three individual parsers allowed the simultaneous execution of the APIs. Simultaneous execution provided the benefit of recording multiple tweets simultaneously. The three parsing programs were created used an API based on a username, a hashtag, and a Twitter query. Simultaneous execution increased the rate of data collection.

### **3. Key Word Generation**

The parser created for this research (using the Twitter query API) also used key word generation. Using a key word generator, the key words for the query pertaining to the object of interest were selected. Using the key words in conjunction with the Twitter search API, converted the resulting tweets to the Twitter search to an RSS feed. The parser was then rewritten based on the XML tags used in this RSS feed to parse the information.

## **B. PROCESSING THE DATA**

Recording a feed and storing it in a database does not imply that everything recorded correlates. Tweets and retweets recorded contained confounding data that was outside the scope of this thesis, i.e., HTTP links and images. The intent of this thesis was to process usable information in Twitter that applied. Although HTTP links and images can be processed for detailed information, i.e., metadata, the level of application programming to retrieve the information pertaining to the locating a person of interest was outside the boundaries of this research. Records containing any form of HTTP link or image were considered useless (spam). A filter was created using the features in SQL to remove any record that contained spam. The records that contained spam were filtered out of the dataset in which it was found and written to a new table

for possible future use. The rule set for spam was any data that contained an HTTP link to information outside of Twitter or an image.

The remainder of the data was separated and sorted in the following manner. Datasets were created by clustering the data with like usernames, hashtags, or generated key words using SQL wildcards. The datasets were then ordered by the associated date-time data. This resulted in multiple datasets that were chronological order pertaining to a specific Twitter handle, hashtag, or key word. There were datasets that overlapped because they shared usernames, hashtags, or key words. The datasets were then analyzed for correlations.

### **C. CORRELATION PROCESS**

The Twitter format, support structures, and procedures, aided the manual process for correlation in two ways: 1) retweets contain data that relate them to the original tweet regardless of the number of retweets and 2) allowing the use of hashtags. Software tools (some automated, some manually enabled) were used to correlate the retweets to tweets. Twitter established a way for users to correlate tweets analyzing the data, i.e., the username of the original tweet, linking the original tweet to the retweet. Likewise, correlation of tweets based on hashtags was performed by analyzing the tweet for the use of a hashtag “#.” When a user establishes a hashtag, the hashtag is recorded in a separate database controlled by Twitter. Once a hashtag is created, any user that enters the same hashtag in a tweet causes that tweet being keyed to be related to every other tweet using the hashtag. This linkage automatically links and correlates tweets based solely on the hashtag, regardless of the rest of the data contained in the tweet. The Twitter API made it possible to record every tweet that uses a hashtag; however, the hashtag had to be known before the API can be used. The same process was used for retweets that contained hashtags, but tweets and retweets are differentiated by Twitters use of the prefix “RT.” The process becomes more difficult because retweets are linked to the original tweet even though it may contain irrelevant information.

The difficult process was when correlating tweets that did not have associated retweets or did not contain a hashtag. Using the source data, information about the tweet was extracted such as the username, geolocation, and the time of the tweet. This information was analyzed and used to correlate the tweet with others in the database.

The data that remained for processing was the tweet itself; the tweeted text can be processed for correlation. If a correlation has not been determined in the previous steps, using the text within a tweet can be used after it is extracted. The current method to correlate tweets based on the tweet itself was through human analysis. To reduce the time required to complete this process, the key word generation approach used by advertisers on search engines was applied. It is assumed a tweet contains a key word. The datasets were then sorted and searched for relationships based on the key word. The existence of a relationship results in a correlation point. A key word may become less needed because most users are currently marking their tweets with hashtags. However, key word generation may prove to be essential in the future as either a facilitator of correlation or a surrogate (since there was no guarantee that a tweet contained a hashtag or retweet). Identifying and using a key word was necessary (and sufficient) to facilitate the correlation process for the purpose of this research.

#### **D. MONITORING**

Correlated reports can be analyzed further through a process of monitoring for key words, statistical deviations from posed patterns, and temporal domain analysis of the time-sequenced tweets. It was beyond the scope of this thesis to analyze tweets for every attribute of society that was contained or reflected in the data streams. However, it was noted in the previous section and Appendix E that the correlation and triangular methodology used in this thesis proved useful in the pilot studies carried out in this research to identify the areas of focus and scope of the work. For future work, identifying and following the attributes of society may provide additional information that will aid in locating

and tracking persons of interest. Additionally, there may be other indication of societal metrics that are also visible in the social network data. Monitoring the frequency of correlated tweets, a pattern can be developed and recognized. Monitoring Twitter during a 24-hour period without any particularly noteworthy events that focus the attention of a large group of users, a frequency of postings can be found and set as a base frequency. The base frequency is the first recognized pattern, from which analysis and correlation can then proceed.

This research shows that when the pattern changes under known conditional changes, it is reflected in the changing frequency of postings. Knowing the conditional changes that are in effect, a new pattern can be recognized. Various patterns were observed that aided in geolocating a person of interest, as well as other patterns that seemed to correlate with movements of users and persons of interest. As the frequency of tweeting changes, a second pattern is recognizable. The more that is known of the circumstances and context of the user, the more recognizable patterns can be found. The patterns are reflective of the boundaries and constraints placed on the domain of correlation and integration.

This research monitored Twitter for several weeks, with focus on a three correlation points resulting in a few hundred thousand records. The focus points for correlation were: 1) Mitt Romney, a high-profile candidate with high visibility, 2) Justin Bieber, a Canadian pop singer-songwriter, and 3) Katy Perry, an American singer-songwriter. Because they had high visibility, their names were found on Twitter along with many other websites and news media, which provided the knowledge of conditional changes. Additionally, they had several websites dedicated to them, which were used to confirm their locations determined in this research.



THIS PAGE INTENTIONALLY LEFT BLANK

## **VI. ANALYSIS**

### **A. INITIAL ANALYSIS**

The method developed to retrieve information from Twitter was implemented to record the tweets and retweets associated with three high-profile individuals. The method resulted in nearly 300,000 public tweets being recorded. Each recorded tweet consisted of the following data: the date, the time, the username, the Twitter GUID and the tweet [Appendix C]. The SQL filter was executed to remove spam tweets leaving about one-half of 300,000 tweets containing usable data. The filtered tweets contained data (such as HTTP links and images) that was outside the scope of this research, but were saved for possible future use. The remaining data was ordered chronologically. The frequency was determined to be 0.146 tweets per second (TPS).

Smaller datasets were created by clustering the remaining data according to the person of interest. Analyzing a single dataset at a time, the data was ordered chronologically and the average frequency was determined. The process was repeated for the remaining persons of interest. Analysis showed the average frequency of tweets and retweets related to the individual persons of interest are unique. The frequency of tweets and retweets related to Mitt Romney, Justin Bieber, and Katy Perry were 0.176TPS, 0.221TPS, and 0.130TPS, respectively.

The frequency for a person of interest is statistically different from others, but the correlation process is the same. Plotting the frequencies for comparison made it clear that each person of interest has a unique Twitter frequency. It also showed that the tweet frequency of the individuals was independent of the whole, i.e., the frequency of all the data is statistically different from the individuals and the individual frequencies are statistically different from each other's. The statistical difference between the individuals shows the frequencies for each are independent and unique. Repeating the process resulted in a method that used the repeated pattern to facilitate finding a person of interest in the social network.

The principal mechanism of correlation was based on the correlation of the data and identifying the associated frequency.

## B. IDENTIFYING CHANGES

The average frequency found in the initial analysis for the person of interest was set as a base frequency, from which a change in frequency can be identified. Figure 7 shows the tweet frequency (calculated in intervals of twenty-minutes) for the three persons of interest over a 24-hour period. It can be seen that each plotted frequency varies throughout the day. The changes in frequency throughout the day caused the base frequency to change slightly because it is inclusive of the new data. As more data is averaged in, the base frequency remained relatively constant. Even though the base frequency changed slightly, it is clear when there is a change in the tweet frequency that draws a closer look.

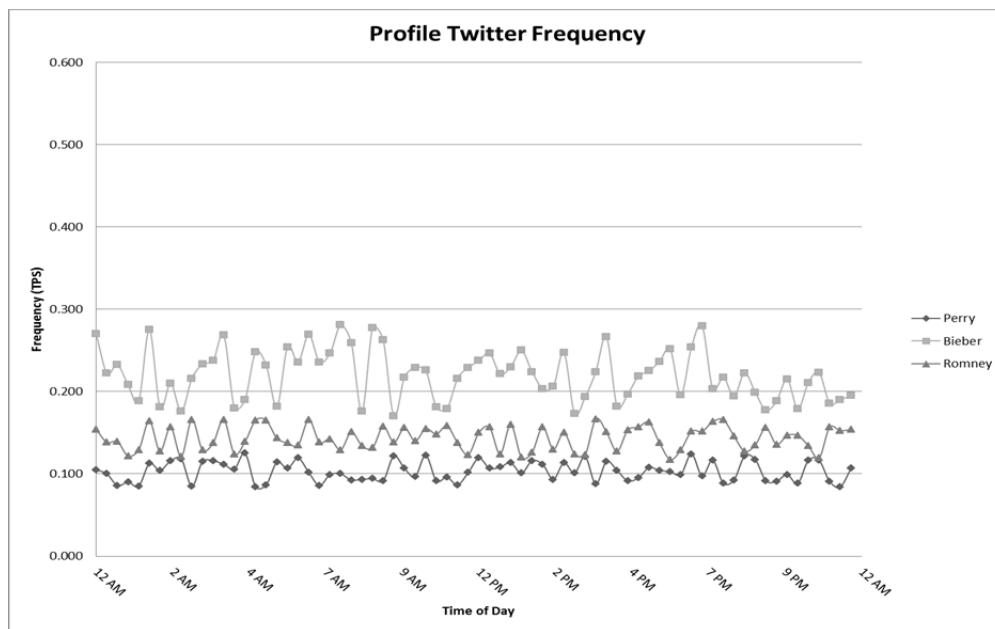


Figure 7. Frequency Comparison of Three Twitter Profiles

## C. CONCENTRATED ANALYSIS

Further analysis of the data before and after the change in frequency occurred revealed a common factor in all three datasets. The change came after

a new, original tweet was posted. The repercussion of the new tweet caused a frenzy of correlated responses that resulted in a change in the frequency Figure 8. In some instances, the change followed a tweet by the person of interest. In the other cases, it was a tweet regarding the person of interest. It was not determined which source had more of an effect on the change in frequency. Further analysis would be required, but it was determined to be outside the scope of this research.

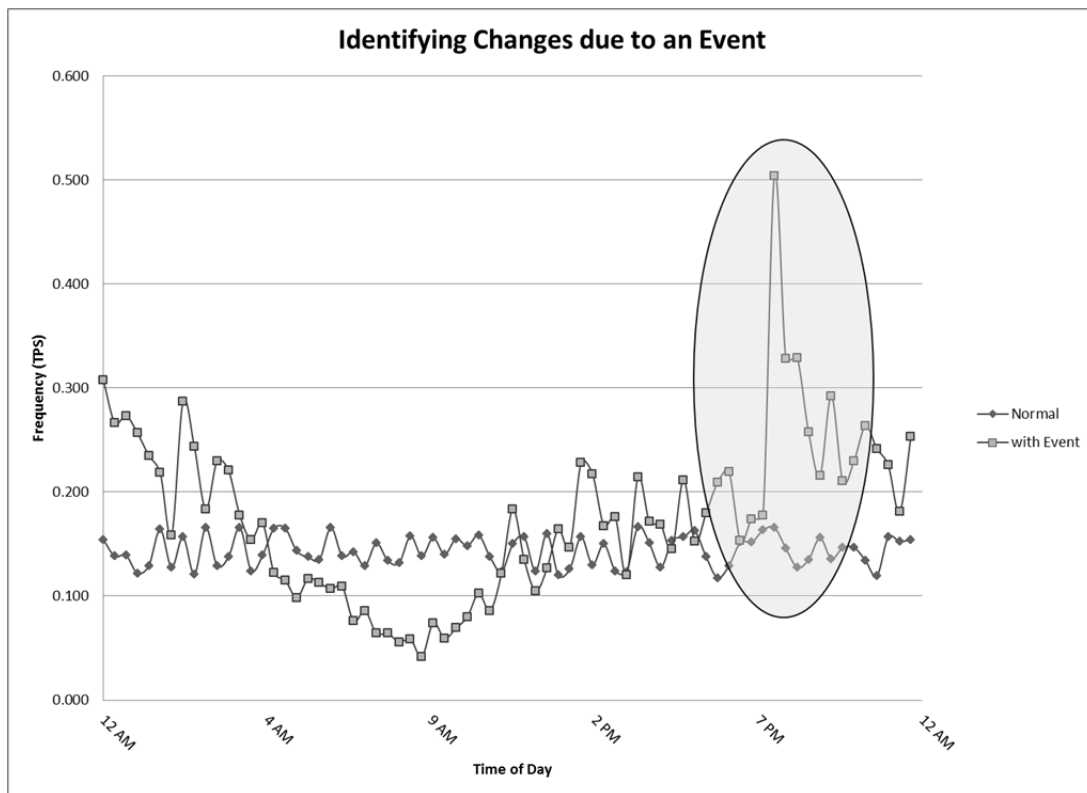


Figure 8. Identifying an Event Through Frequency Comparison

After identifying the cause for the changes in frequency, a cluster of tweets and retweets were identified for concentrated analysis. The clusters were based on the time-line of the change and the tweets that correlated to the change and with the person of interest. The preset amount of time (one hour) preceding the change was selected to be inclusive of the new tweet and possible causes of

the new tweet, which caused the change in frequency. The clusters of data, associated with a change, were analyzed for location information.

#### **D. DETERMINING LOCATION**

Initial analysis revealed a person of interest could be identified by their frequency, and concentrated analysis showed how a change could be identified. There was still a need for more detailed processing of the tweet data to determine a location. Applying the method developed to process the tweet text, additional key words were determined, e.g., the name of an event, organization, or building, and used for correlation. The key words that were associated with Mitt Romney were: NRA, St. Louis, Missouri, address, and convention. The additional key words were used in the SQL query methods to correlate the recorded data based on the determined key words. Again, the associated tweets were ordered chronologically. The usage frequency of the key words was then determined. The associated frequencies were nearly the same for the 5 key words at 0.438 TPS, and they occurred at the same time as the spike in the base frequency for Mitt Romney. Identical analysis was done with the data that was associated with Justin Bieber and Kate Perry. The key words that were associated with Bieber were: Billboard, music, Vegas, date, prom, and Cady. The frequency of use for the key words were nearly the same at 0.395 TPS. For Perry: performance, music, video, and new were the associated key words with a frequency between 0.100 TPS and 0.184 TPS.

The usage frequency of the key words was compared to the base frequency of the person of interest. A correlation was determined to exist if there was an increase in key word usage along the same time-line of the change from base frequency for the person of interest. This was performed by executing SQL query commands on the data combining the key words and the person of interest. Since the data was ordered chronologically, in the cases of Romney, the rapid increase in frequency was found to be directly correlated with the increase usage of the key words. For Bieber, similar results were found. In both cases, by

investigation, there were sufficient key words that correlated to indicate their locations. In the case of Perry, a correlation of key words and a change in frequency was not determined.

Analysis of the correlated data following the determination of a location showed that the whereabouts of a person of interest could be tracked. This was completed by repeating the methods developed to determine the person's location. Because there is a time stamp with every record, after the location was determined and recorded, a location history of the person of interest was created. Each recorded location had an associated time period, but the gaps in time between locations generated a poor tracking tool.

THIS PAGE INTENTIONALLY LEFT BLANK

## **VII. SUMMARY AND CONCLUSIONS**

### **A. SUMMARY**

Twitter contains an immeasurable amount of information in the tweets and retweets published on the social network. There is sufficient information to locate a person of interest. A method was developed and employed to record tweets and retweets. The data embedded in the tweets and retweets was extracted and written to a database for processing. The SQL processes developed in this thesis were implemented to filter, cluster and correlate the data in order to locate the person of interest. Analysis of the data recorded and processed revealed sufficient information to locate a person of interest. This thesis developed a method to extract data from Twitter for the purpose of locating or tracking an object of interest. The method developed identified and recorded the tweets and retweets associated with three high profile individuals on Twitter. In two of the three cases, a correlation was determined and a location could be identified. The step to process the data and determine sufficient information to track an individual was developed; however, analysis of the correlated data showed gaps in time that lead to a poor method of tracking a person of interest.

The information on the location of a person of interest is readily identifiable in approximately one-third of single tweets by human analysis. The methodology and tools developed in this research indicates that a person of interest can be located approximately one-half of the time in single tweets. Repetition of the processes developed in this research can be refined and improved upon for other applications such as HA/DR information gathering or for emergency response.

### **B. CONCLUSION**

Extracting information from a social network is problematic and confounded by access to the data and the type of data without the appropriate set of tools. The scope of this thesis was to investigate the technology issues



relating to the processing and analysis of Twitter data, recognizing that the confounding factors may be more significant than the actual algorithm that was used to correlate the data. This research introduced a method that defined and provided the tools to utilize a social network to monitor various activities reported by the Twitter users. This thesis focused on finding the location of a person of interest. Specifically, the frequency of tweets and retweets are a direct indication of the activity of an object of interest from which correlations between the data can be made. The objective was to use the data which was correlated as the source of data to determine the location.

Using the Twitter APIs, the confounding factor of access to the data was mitigated, and the information found in Twitter was recorded in a database. The SQL filter developed in this thesis was used to lessen the confounding issue found in the type of data shared in Twitter. Analysis of the data supported the key assumption that people are communicative about themselves and others. The data was clustered into datasets that related to three persons of interest. Ordering the data chronologically, made it possible to determine the frequency of Twitter usage that related to the persons of interest and the frequency of key words. The frequency of tweets and retweets related to Mitt Romney, Justin Bieber, and Katy Perry were 0.176TPS, 0.221TPS, and 0.130TPS, respectively. Utilizing the method developed in this thesis, the frequency of key words was correlated to the frequency associated with the person of interest. The frequencies associated with key words in the Romney dataset averaged 0.438 TPS, which occurred at the same time as the spike in the frequency of tweets associated with Romney. The key words for Bieber were: Billboard, music, Vegas, date, prom, and Cady averaged a frequency of 0.395 TPS. Like Romney, the frequency spike for the key words occurred at the same time as the spike in the frequency of tweets associated with Bieber. For Perry: performance, music, video, and new were the associated key words with a frequency between 0.100 TPS and 0.184 TPS, but a spike in tweet frequency could not be identified. Analysis of the correlated data revealed it was possible to determine the location

of two out of three persons of interest. The location of Romney was St. Louis, MO, and the location of Bieber was Las Vegas, NV. The location of Perry could not be determined because there was no identifiable spike in the frequency of the associated data.

THIS PAGE INTENTIONALLY LEFT BLANK

## **APPENDIX A: A TWITTER CASE STUDY**

The following case study was done in March 2010 by Kaitlin LaCasse and Laura Quinn, employees of Idealware. Idealware is a 501(c)(3) nonprofit that helps nonprofits make smart software decision by providing resources about software. There are two important takeaways from the case study. The first item to note is the emphasis on the unique culture of Twitter, which revolves around sharing resources, but the difficult involved without an appropriate tool. The second item of importance is the confirmation that Twitter can be an effective tool in communications. The study can be found at: <http://idealware.org/articles/reaching-out-wide-audience-twitter-case-study1>.

...

### **REACHING OUT TO A WIDE AUDIENCE: A TWITTER CASE STUDY**

By Kaitlin LaCasse and Laura Quinn, March, 2010

Is Twitter useful for nonprofits? It's certainly been useful for Idealware. In this detailed case study, we talk through how we use Twitter, the results we've seen, and how it might be helpful for your organization.

Twitter has become a hot topic these days. Companies, nonprofits and government agencies are scrambling to put Twitter strategies in place, and businesses, consultants and bloggers alike are "tweeting out" updates on everything from marketing tips to the state of their children's potty training.

Despite its popularity, it can be difficult for the uninitiated to understand how Twitter can be useful. OK, so you can send out 140-character messages to people who follow you. Isn't that similar to what you can do with Facebook, or e-mail, or a blog—with the extra difficulties of length restrictions? Well, yes. But the

unique culture of Twitter, which revolves around passing resources on to others, makes it really quite useful as an outreach and communications channel.

It could be an effective addition to your nonprofit's current communication mix. It certainly has been for Idealware. We weren't really believers ourselves when we first started trying it out last summer, but it's proven to be a useful outreach tool that takes less time than a lot of other social media tools. This article walks through what we do with Twitter, and the lessons we learned along the way.

## **GETTING STARTED**

Getting started with Twitter is easy. Simply go to [www.twitter.com](http://www.twitter.com) and choose a user name. In fact, we found it a little too easy. Over the summer of 2009, we set up an Idealware account to see what Twitter was all about, with the intention of "listening" to what people were saying and getting a sense for the culture, without jumping in quite yet. Unfortunately, a number of people saw that we'd registered and started publicly lobbying, through Twitter, to "make Idealware tweet." In the face of public pressure, we didn't have much choice but to start tweeting. Lesson learned: you might want to start with a personal account or an alias before you join as a well-known individual or organization.

We use "Idealware" as our username, or "handle." On Twitter, a username is frequently shown with the @ sign, so we're @Idealware. tweeting as an organization can be awkward at times, as the person-to-person culture of Twitter is geared more toward personal accounts. There are two different people tweeting on our account, which we make clear in our Twitter bio, but there's no obvious way for followers to tell which of us posted a given tweet.

There's a negative side to tweeting under an organizational name. It's hard for people to feel like they're having a conversation with a logo. On the other hand, it

makes it easier for people to share responsibility for the account and maintain a steady stream of tweets. It also keeps one person from becoming the “face” of the organization, which could be a positive or negative depending on your goals.

What are your options, other than using an organizational handle?

- Use your own name. For example, Wendy Harman, the Red Cross’s social media guru, tweets from @wharman. It’s fairly strictly a professional account (as opposed to including information on her social life), and she includes the Red Cross in her bio. This obviously puts more of a personal touch on the tweets, but it may not be immediately clear that you’re representing your organization.
- Use some combination of your name and your organization. Holly Ross, Executive Director of NTEN, tweets under @ntenhross. Her handle identifies her as an individual, using Twitter for professional purposes. In this way it’s clear who is tweeting, and that she is doing so on behalf of the organization. On the other hand, with this method it’s hard to share the work among multiple people.

You can’t change your Twitter handle once you’ve created it, so it’s important to think it through with care.

Once we set up an account, we chose a picture to be associated with the account, also called an “avatar.” This picture has to be square, which poses some issues when using a horizontal logo like ours, so we went with an image of the letter “I”... with which we’re only somewhat satisfied. The constraints are more accommodating for a picture of a person, or perhaps an image that represents your mission (for example, Charity:Water uses a picture of a child drinking water). You can change your avatar at any time.

Our profile also includes a bio that gives our location, a link to our website and a

short description of what we do. Twitter bios tend to be less formal than what you might have on your website, and show a little bit of personality.

## **LISTENING**

Despite our all-too-public introduction to Twitter, we still wanted to start mostly by listening to what people were saying about topics of interest to us. How did we find people and conversations? By searching for key terms, like “social media” and “nonprofits,” that were relevant to our organization.

We searched—and continue to search—Twitter “hashtags” in order to see who is talking about things of interest. A hashtag is simply a keyword with a pound symbol (#) or “hash sign” in front of it. People use hashtags to mark their tweets, which makes it easy to search by topic. For example, we follow the hashtag #nptech.

As we learned who was saying things that were useful for us, we began to “follow” those people—basically, subscribing to their Twitter feed—to make it easy to see everything they post. These people often pointed us toward other people worth following. In addition, we follow people who are part of the core Idealware community, like our bloggers and board members. You can find people on Twitter by searching for them, or you can ask your contacts for their Twitter handles.

## **JOINING THE CONVERSATION**

As we felt more comfortable with Twitter, we started tweeting more. What do we post?

- **Retweets:** One of the most common things to do on Twitter is to “retweet,” or RT, a tweet relevant to your community. As our mission is to provide information to help nonprofits make smart software decisions, passing along resources that our audience might find helpful is a great adjunct to other information we provide.
- **Posting links:** Similarly, if we come across a resource elsewhere we’ll often post it via Twitter. We use <http://bit.ly> to automatically shorten long link addresses into much shorter ones that will better fit within the 140-character limit. Bit.ly also allows you to track the number of people who click on a link.
- **New articles, reports, and blog posts:** When we have a new article, report, or blog post, we tweet about those. For a big report, we might post a couple of times – for instance, to announce that it’s coming soon, then providing a link to the live report, then thanking people for retweeting it.
- **News and promotions:** If something’s happening that we’re hoping to spread the word about—for instance, a particular event or a program we’re trying to recruit people for—we’ll tweet about it.
- **“Behind the scenes” information:** We regularly tweet information about what’s going on at Idealware, or about what we’re working on—for example, we tweeted that we were putting an upcoming report into layout. People responded that they couldn’t wait to see it, so it not only gave an insiders-view into our work, but it helped to increase excitement about an upcoming project. This is an advantage of Twitter— it’s straightforward to keep people up-to-date on a project on even an hour-by-hour basis... which also keeps it at the top of your supporters’ minds.
- **Questions:** If we ask a question, people will often respond. Often the actual questions relate to fairly obscure areas of software, and the quality and number of answers we receive depend on who happens to be on Twitter at the time. This is a great way for us to find examples and case studies of commonly used software, though.



- **Responses to other people’s questions or comments:** It’s always great when you can start a relevant conversation.
- **Thanking people for retweeting our posts:** People often publically thank others, using their handles, for retweets. We have kind of mixed feelings about this—it can generate a lot of noise on your account if a lot of people are retweeting you, but it does act as a nice thank you, as it could introduce your whole follower base to that retweeter’s account. It’s also a great way to keep saying things about something that you really want to promote the heck out of—for example, you can keep thanking people for retweeting info about a report... and keep providing the URL for other people to find the report.

In general, we try to post a couple of times a day. This takes about two-to-three hours a week to keep up with Twitter and remember to post. Unlike some other channels, a lot of volume is completely acceptable on Twitter—we could likely tweet up to 10 times a day or so if we had useful things to say (and had the time). Going a couple of days without posting is likely to make people wonder what happened to you.

One thing we tried that didn’t work so well was a “Factoid” tweet. If we were doing research, we might tweet something interesting that we just learned. We have no idea if people found them helpful, but we never got any sort of response, and they were rarely retweeted.

Note that Twitter is a public venue. Everything we post can be seen by anyone who has chosen to follow us. We can send someone a “direct message”—a tweet that goes only to them—but we can’t send a tweet to a select group of people or tailor a different message to target audiences. It’s also worth noting that tweets come up in Google searches.

## **BUILDING OUR FOLLOWER BASE**

We've gone from a brand new account to having around 1,350 followers in about eight months—interestingly, we have about four times as many Twitter followers than Facebook fans, though we started both in a similar timeframe. We haven't done any substantial promotion—rather, it's been more of an organic process that incorporated the following steps:

- **Following people.** We found it really important to start by following a number of people. In Twitter, most people will follow you if you follow them—it's considered proper Twitter etiquette. This is a great way to seed your follower base with people you think might be actually interested in what you have to say.
- **Saying interesting things.** For us, the most useful way to gain new followers is to tweet something that's widely retweeted. When someone retweets our post, all of their followers see that we exist—and some decide to follow us. Our articles and report announcements get the most retweets by far.
- **Retweeting people.** Retweeting other people is also a great way to get on their radar, and to encourage them to follow you.
- **Promoting through other channels.** We've actually done very little promotion of our Twitter account to those who follow us primarily via e-mail or our blog. There's a link to it on our website, and in the standard footers on our e-mails, but that's about it. We would likely gain some more followers by doing more cross-promotion.

We're seeing steady, perhaps even exponential, growth over time, and a lot of retweets by prominent entities, like publications and prominent bloggers.

## **MANAGING TWITTER**

Twitter itself does not have a very user-friendly interface, leading several vendors to develop other ways to interact with Twitter. We use a tool called tweetDeck—

truthfully, it's hard to imagine being able to effectively keep up our Twitter account without it, or something like it.

You download tweetDeck directly onto your desktop, where it automatically alerts you to new tweets (with Twitter, you have to refresh your browser) and allows you to monitor different groups on a single screen. For example, we monitor any mention of the term “Idealware” on Twitter, and can see that in a column in the tweetDeck interface.

We also filter down the list of people we are following to create a list of people who generally tweet things of interest to us. If you're going to follow everyone who follows you, as we do, this is a critical and common step to cut through the clutter. So how many of our “followers” are actually paying any attention to us? It's impossible to tell, except by measuring who actually responds or retweets.

HootSuite is another common option. HootSuite is an online tool with a multi-column similar view to tweetDeck. HootSuite also lets you schedule tweets for the future, which might be helpful for a targeted campaign. You can also have multiple accounts in one interface, and multiple “editors”—essentially, all the people you want to have tweeting from the same Twitter handle.

## **WHAT'S THE BANG FOR THE BUCK?**

So is Twitter working for us? We think it is, in a number of ways:

- It's a great way to quickly spread the word about a particular resource to a much wider audience than our own list. For instance, we recently posted about the results of our Social Media Survey. It was widely picked up and retweeted by dozens of people, meaning it went out to all of their personal networks, including prominent sources like Harvard Social Media and the Foundation Center. It's still circulating on Twitter as I write this, about a

week later. It's hard to track the exact number of click-throughs to actually view the report, but it's more than 400 people just from Twitter alone.

- It's very useful for learning about new resources, and what partners and potential partners are up to—in general, just to monitor what's happening in the community that we work in.
- It's useful to reach a more-targeted audience than our e-mail list. In general, those following us on Twitter are more likely to be tech-savvy partner organizations and consultants than those on our e-mail list. This makes it a good place to promote opportunities geared toward these audiences (like licensing options and sponsorship opportunities). In fact, we made contact with a new client through Twitter with whom we did a substantial amount of paid work.

What's the investment? All the tools involved are free. It comes down to the time invested. It takes about two-to-three hours a week for us to do the bare minimum of our strategy—keeping somewhat on top of what others are saying, and posting a few times a day. This doesn't include actually reading the relevant resources that other people have posted. You could easily spend more time.

A couple of times a day, Kaitlin, who was our Americorp VISTA social media specialist, scanned through what was posted, saw if we need to respond to anything, and retweeted or posted things of interest. Laura, our Executive Director, was swooping in in a bit more sporadically to post something or see what people are saying -- she's now doing most of the posting, with the help of Andrea Berry, our development director. We find that it doesn't take a lot of coordination between the staff, unless we're trying to widely promote a particular resource with care. It is easy, however, to fall down the Twitter rabbit-hole—sometimes, when we need to get work done, we both need to just shut down

Twitter to concentrate.

Overall, we definitely find it a useful investment for Idealware, and think it has real possibilities for a number of nonprofits. It likely will come down to your audience and your organization, though. Are there people you'd like to communicate with on Twitter, which tends to be a community of working-age tech- and media-savvy folks? Do you want to get the word out about things that are straightforward to share on Twitter?

If so, we think Twitter's worth a try. Define what you hope to achieve with it, start by listening, and then dive right in!

License: [Creative Commons Attribution-NonCommercial-NoDerivs 2.5](#)

1. K. LaCasse and L. Quinn, Idealware, "Reaching Out to a Wide Audience: A Twitter Case Study," March 2010, <http://idealware.org/articles/reaching-out-wide-audience-twitter-case-study1>.

## **APPENDIX B: JOPLIN TORNADO CASE STUDY**

The following case study was initiated by Dave Bourne in May 2011. As the Corporate Communications Manager for the Scarborough Hospital (Toronto, Canada), he raised the question of integrating social media into the hospitals strategy for crisis communication following the tornado that occurred May 22, 2011, in Joplin, Missouri. Dave focused on how the staff of St. John's Medical Center Joplin used social media to direct and inform family members looking for the location or information pertaining to those injured in the tornado. Responses to Dave's question complete the case study, and as follow up to his case study, Dave developed an index to measure the reputation of his hospital in Toronto. His presentation titled "Taking the pulse of healthcare social media: A prescription for measuring reputation" is important to look at because it provides an example how the interaction of people in social networks can have an influence. The case study can be obtained from <http://www.smich.ca/?p=297> and Dave's follow on index presentation can be found at [http://www.slideshare.net/d\\_bourne/taking-the-pulse-of-healthcare-social-media](http://www.slideshare.net/d_bourne/taking-the-pulse-of-healthcare-social-media).

### **Missouri tornado provides a case study in social media crisis communication**

A few weeks ago, I posted a topic for the #hcsmdca chat about using social media as part of a hospital crisis communication plan. I was interested in finding out whether any hospitals had formally integrated social media into their strategies for communicating during a crisis.


While many of us are using Twitter and Facebook regularly, I wondered whether anyone had put any thought into leveraging these platforms to help manage the information demands that can tap our resources during times of crisis.

Just this weekend, we saw an example of a hospital facing a major crisis, and using social media to engage with the general public, families of victims, volunteers and the media.


In Joplin, Missouri, the St. John's Medical Center took a direct hit from the massive tornado that devastated the community. At least 116 people were reportedly killed by the twister, including five patients at SJMC.

During the chaotic hours following the tornado, SJMC lost their website, greatly limiting their ability to communicate with the outside world. As media queries poured in, along with frantic pleas from family members trying to locate patients, Facebook became the communication tool of choice.

Wall posts show how the hospital's staff — it is unclear whether they were acting in an official capacity or not — directed family members to special hotline numbers. This was in direct response to the many posts from people searching for loved ones. The Facebook page also helped steward anyone interesting in donating or volunteering, greatly reducing the potential strain on the hospital's phone system.

There were few Twitter [@StJohnsHealth] updates from the hospital, with the exception of some early reports about patients being transferred to neighboring hospitals, and some later posts redirecting to the Facebook page.

How do you think St. John's Medical Center fared using social media in such a trying circumstance? The potential loss of Internet access when an entire town is destroyed by a disaster is something that many of us may not have considered in our own crisis communication plans. Could SJMC have done better? What lessons can we all learn from this case study?

Dave Bourne,  Manager of Corporate Communications

The Scarborough Hospital

@d\_bourne

@ScarboroughHosp 

dbourne@tsh.to

## COMMENTS TO “MISSOURI TORNADO PROVIDES A CASE STUDY IN SOCIAL MEDIA CRISIS COMMUNICATION”



**ann.fuller** 25 May 2011 at 9:23 am #

I think it's a great question Dave.

We don't currently have it built into our crisis communications plan. But it really has me thinking both about how we could use social media to communicate with patients and families — and also with staff. It can be equally difficult to reach staff in times of crisis.

Has anyone seen anything done from that perspective?

**Reply**



**Whitney** 25 May 2011 at 11:52 am #

It is obvious that technology can assist hospitals and medical centers in a multitude of ways and can be noted within the above examples. The sad reality is that although a lot of hospitals have websites and utilize social media, a multitude of medical facilities do not. These facilities can face the same issues, limiting their ability to communicate effectively especially during times of crisis. I hope that all health care facilities will eventually utilize and implement the necessary technology to benefit their services especially during times of crisis.

**Reply**



**@d\_bourne** 25 May 2011 at 4:45 pm #

It's probably safe to assume that most hospitals that are using social media would make at least limited use of these platforms during a crisis. But do you see Facebook or Twitter replacing any traditional crisis communication tactics? In the past, we would primarily be focused on pushing out information.



Now, we can actually engage audiences. Could Twitter or a Facebook page replace daily media briefings, for example, or just enhance the level of communication possible?

And what about having followers repurpose your messages? In the St. John's Medical Center example, Facebook followers are sharing info about Red Cross donations, volunteer opportunities and hospital phone hotlines, therefore reaching a much broader audience than would have been possible without social media.

To me, this is the real potential of leveraging SM for crisis communications...it's the exponential reach you can achieve with little effort, combined with the ability to facilitate two-way conversations.

## APPENDIX C: MLB TWEET

### A. MLB TWEET IN TWITTER FORMAT

Figure 9 is a copy of a tweet posted on Twitter by Twitter handle @MLB and can be found at: <https://twitter.com/MLB/status/207905572677353473>.



Figure 9. A tweet posted by Twitter handle @MLB at 11:45 am on 30 May 2012

### B. SOURCE CODE

The following is a copy of the source code of the same, single tweet submitted by Twitter handle @MLB in original HTML format:

```
<i class="dogear"></i>

<div class="content">

  <div class="stream-item-header">
    <small class="time">
      <a href="/#!/MLB/status/207905572677353473" class="tweet-timestamp js-permalink" title=""><span class="_timestamp js-short-timestamp" data-time="Wed May 30 18:45:25 +0000 2012" data-long-form="true"></span></a>
    </small>
    <a class="account-group js-account-group js-action-profile js-user-profile-link" href="/MLB" data-user-id="18479513">
      
      <strong class="fullname js-action-profile-name show-popup-with-id">MLB</strong>
      <span>&rlm;</span><span class="username js-action-profile-name"><s>@</s><b>MLB</b></span>
    </a>
  </div>

  <p class="js-tweet-text">
```

Angels in the outfield: Trout, Bourjos made heavenly plays vs. the Yankees last night: <a data-expanded-url="http://atmlb.com/JMLXJn" class="twitter-timeline-link" href="http://t.co/Wr0xZHM9" rel="nofollow" class="twitter-timeline-link">atmlb.com/JMLXJn</a>  
</p>

<div class="stream-item-footer">

<div class="context">  
</div>

<a class="details with-icn js-details" href="/#!/MLB/status/207905572677353473">  
<span class="details-icon js-icon-container">  
</span>  
<b>  
<span class="expand-stream-item js-view-details">  
  
<span class="expand-action-wrapper">  
Expand  
</span>  
</span>  
<span class="collapse-stream-item js-hide-details">  
  
</span>  
</b>  
</a>

<ul class="tweet-actions js-actions">  
<li class="action-reply-container">  
<a class="with-icn js-action-reply" data-modal="tweet-reply" href="#"  
title="Reply">

<i class="sm-reply"></i>  
<b>Reply</b>  
</a>  
</li>  
<li class="action-rt-container">  
<a class="with-icn js-toggle-rt" data-modal="tweet-retweet" href="#">  
<i class="sm-rt"></i>

<b><span class="undo-retweet" title="Undo retweet">Retweeted</span><span  
class="retweet" title="Retweet">Retweet</span></b>  
</a>

</li>  
<li class="action-del-container">  
<a class="with-icn js-action-del" href="#" title="Delete">  
<i class="sm-trash"></i>  
<b>Delete</b>  
</a>  
</li>  
<li class="action-fav-container">  
<a class="with-icn js-toggle-fav" href="#">  
<i class="sm-fav"></i>

```

        <b><span class="unfavorite" title="Undo favorite">Favorited</span><span
class="favorite" title="Favorite">Favorite</span></b>
    </a>
</li>
</ul>    </div>

    <div class="expanded-content js-tweet-details-dropdown">
    </div>
</div>
</div>

<div class="js-stream-item stream-item stream-item" data-item-
id="207896460082159621-promoted" data-item-type="tweet" id="stream-item-tweet-
207896460082159621-promoted">

    <div class="tweet original-tweet js-stream-tweet js-actionable-tweet js-hover js-profile-
popup-actionable js-original-tweet
promoted-tweet

“

data-tweet-id="207896460082159621"
data-impression-id="4f8a27f0fd271a09"
data-item-id="207896460082159621"

data-screen-name="NewsGator" data-user-id="14186545"

data-promoted="true"

data-is-reply-to="false"
data-
status="{&quot;in_reply_to_status_id_str&quot;:null,&quot;possibly_sensitive&quot;:false,&quot;id
_str&quot;:&quot;207896460082159621&quot;,&quot;contributors&quot;:null,&quot;in_reply_to_u
ser_id&quot;:null,&quot;in_reply_to_status_id&quot;:null,&quot;in_reply_to_user_id_str&quot;:null
,&quot;retweeted&quot;:false,&quot;created_at&quot;:&quot;Wed May 30 18:09:13 +0000
2012&quot;,&quot;user&quot;:{&quot;id&quot;:14186545,&quot;screen_name&quot;:&quot;News
Gator&quot;,&quot;time_zone&quot;:&quot;Mountain Time (U.S. &
Canada)&quot;,&quot;statuses_count&quot;:1816,&quot;id_str&quot;:&quot;14186545&quot;,&qu
ot;profile_use_background_image&quot;:true,&quot;profile_text_color&quot;:&quot;444444&quot;
,&quot;is_translator&quot;:false,&quot;location&quot;:&quot;Denver,
CO&quot;,&quot;following&quot;:false,&quot;utc_offset&quot;:-
25200,&quot;profile_sidebar_border_color&quot;:&quot;444444&quot;,&quot;name&quot;:&quot;
NewsGator&quot;,&quot;default_profile_image&quot;:false,&quot;notifications&quot;:false,&quot;f
avourites_count&quot;:2,&quot;protected&quot;:false,&quot;profile_background_tile&quot;:false,&
quot;contributors_enabled&quot;:true,&quot;friends_count&quot;:1461,&quot;profile_sidebar_fill_
color&quot;:&quot;dcddde&quot;,&quot;geo_enabled&quot;:false,&quot;profile_background_imag
e_url_https&quot;:&quot;https://si0.twimg.com/profile_background_images/326641794/Twitter
BG9.11.png&quot;,&quot;description&quot;:&quot;Bringing Enterprise and Government 2.0 Social
Computing features directly into Microsoft
SharePoint.&quot;,&quot;show_all_inline_media&quot;:false,&quot;follow_request_sent&quot;:fal
se,&quot;verified&quot;:true,&quot;profile_background_color&quot;:&quot;444444&quot;,&quot;pr
ofile_image_url_https&quot;:&quot;https://si0.twimg.com/profile_images/1017911543/Activity-
Stream_normal.png&quot;,&quot;default_profile&quot;:false,&quot;profile_image_url&quot;:&quot;
http://va0.twimg.com/profile_images/1017911543/Activity-

```

Stream\_normal.png",&quot;followers\_count":3110,&quot;lang":&quot;en",&quot;url":&quot;http://www.newsgator.com",&quot;profile\_background\_image\_url":&quot;http://a0.twimg.com/profile\_background\_images/326641794/TwitterBG9.11.png"ot,&quot;created\_at":&quot;Thu Mar 20 19:54:07 +0000 2008",&quot;profile\_link\_color":&quot;76b142",&quot;listed\_count":275,&quot;truncated":false,&quot;in\_reply\_to\_screen\_name":null,&quot;entities":{&quot;ot;user\_mentions":[{&quot;name":&quot;Microsoft",&quot;id\_str":&quot;74286565",&quot;indices":[27,37],&quot;screen\_name":&quot;Microsoft",&quot;t;id":&quot;74286565"}],&quot;urls":[{&quot;display\_url":&quot;bit.ly/IF2Azu",&quot;t;expanded\_url":&quot;http://bit.ly/IF2Azu",&quot;indices":[102,122],&quot;url":&quot;http://t.co/vfZGECI0u"}],&quot;hashtags":[{&quot;indices":[87,94],&quot;text":&quot;socbiz"}],&quot;source":&quot;web",&quot;place":null,&quot;retweet\_count":0,&quot;favorited":false,&quot;id":207896460082159621,&quot;coordinates":null,&quot;geo":null,&quot;text":&quot;See how NewsGator enhanced @Microsoft\u2019s Student Partner Portal to harness the power of #socbiz tools: http://t.co/vfZGECI0u",&quot;promoted\_content":{&quot;impression\_id":&quot;4f8a27f0fd271a09",&quot;disclosure\_text":&quot;&quot;,&quot;disclosure\_type":&quot;&quot;,&quot;promoted",&quot;social\_context":[]}}"

data-expanded-footer=""

### C. SOURCE CODE IN XML FORMAT

The following is a copy of the source code of the same, single tweet submitted by Twitter handle @MLB in XML format:

```
<item>
  <title>MLB: Angels in the outfield: Trout, Bourjos made
heavenly plays vs. the Yankees last night: http://t.co/Wr0xZHM9</title>
  <description>MLB: Angels in the outfield: Trout, Bourjos made
heavenly plays vs. the Yankees last night:
http://t.co/Wr0xZHM9</description>
  <pubDate>Wed, 30 May 2012 18:45:25 +0000</pubDate>

<guid>http://twitter.com/MLB/statuses/207905572677353473</guid>

<link>http://twitter.com/MLB/statuses/207905572677353473</link>
  <twitter:source>&lt;a
href=&quot;http://www.awarenessnetworks.com/home/&quot;
rel=&quot;nofollow&quot;&gt;Social Marketing
Hub&lt;/a&gt;</twitter:source>
  <twitter:place/>
</item>
```

## **APPENDIX D: EXCERPTS ON CASE STUDIES**

### **A. EXCERPTS ON CASE STUDY RESEARCH FROM LANGFORD 2012A AND LANGFORD 2012B**

A rich source of experiences can be gleaned from previous social situations. The challenge is to extract principles or develop pose a relevant theory by (1) performing some combination of qualitative or quantitative analysis, (2) focusing on ontological structures to organize the contribution, and (3) presenting the concepts, data, and results in a form that is pertinent and leads to understanding and predictability (these comments inspired by Gregor 2006). Building on the idea that a theory explains something based on general principles which are independent of the particulars of that which is to be explained (Oxford Encyclopedic English Dictionary), case study research offers an approach that incorporates structures of validation and reliability (Yin 1981, 1984, 1989) from basic observations of social phenomena. Knowing what you want from the case study frames the manner of investigation and analysis as well as the utility of the results for theory development (Kitchenham and Pickard 1998).

The methods of extracting that information, specifically principles and heuristics is referred to as case study methods. Yin (defines a case study as ‘an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident’ (Yin 2003). By reviewing a single case that is representative of like-kind projects, generalized knowledge can be gained (Flyvbjerg 2006). This knowledge is often in the form of principles and if generalized knowledge is captured by principles, then two key results of an investigative case study are possible and determinable: first, a practical knowledge is present and purposeful, i.e., context dependent knowledge can be generalized and abstracted; and second, through the development of principles and heuristics, general purpose propositions can be stated. Combining practical knowledge that interprets the

scenarios and vignettes of a case study, the theoretic propositions can be assembled and reviewed to glean lessons, coaching, and instruction.

The goal of achieving a predictive capability based on case study methods seems problematic if based merely on the social sciences and their tools. Numerous social science researchers have suggested or stated explicitly that ‘there does not and probably cannot exist predictive theory in social science’ (Campbell 1975; Flyvbjerg 2006). The purpose for apply case study methods in this thesis is to expose the interactions between people and their processes that are undergoing integration at the social level rather than at the deeper phenomenological interactions that result in integration at all levels of systemic interplay.

There is valuable information to be gleaned from case studies – they can be useful for formulating hypotheses, hypotheses testing, theory construction, and developing general theories (Eisenhardt 1989). Eisenhardt’s constructs and research methods are widely used as integrative constructs in systems engineering, business, and sociology, for developing computational theory (Zhongyuan, Rouse, and Serban 2011) for knowledge management theory (Cranfield and Taylor 2008), and for integration frameworks (Themistocleous and Irani 2002), for example. Adapting her eight steps for this thesis and developing coherency and congruence, and applying ‘an appropriate overlay of scholarship’ over the case facts’ (Ferris, Cook, and Honour 2003) , the Eisenhardt approach outlines a roadmap for using case study to identify principles that can be used for correlation. These adaptation of the eight steps transform into (1) defining the research question, (2) selecting a case study that exhibits situations typical of the social scenarios, (3) gathering relevant historical context and qualitative/quantitative data, (4) creating an appropriate method (e.g., Yin 2003), (5) analyzing data using correlative triangulation or like means of relating quantitative and qualitative data, (6) shaping the hypotheses through adduced theoretics, (7) performing a comparative literature survey, and (8) reaching closure on principles by analyzing their scope and validity.

## References:

1. Campbell, D. T. (1975). "Degrees of Freedom and the Case Study," Comparative Political Studies, 8(1), pp. 178–191.
2. Cranfield, D. J. and Taylor, J. (2008). "Knowledge Management and Higher Education: a UK Case Study," The Electronic Journal of Knowledge Management, 6(2), pp. 85–100, available online at [www.ejkm.com](http://www.ejkm.com).
3. Eisenhardt, K. M. (1989). "Building Theories from Case Study Research," The Academy of Management Review 14(4), pp. 532–550.
4. Ferris, T. L. J., Cook, S.C., and Honor, E.C. (2003). A Structure for Systems Engineering Research. Proceedings of SETE 2003, Rydges Capital Hill, Canberra, Australia.
5. Flyvbjerg, B. (2006). "Five Misunderstandings About Case-Study Research," Qualitative Inquiry 12(2), pp. 219–245.
6. Gregor, S. (2006). "The Nature of Theory in Information Systems," Management Information Systems Quarterly 30(1): 62–89.
7. Kitchenham, B., and Pickard, L.M. (1998). "Evaluating Software Engineering Methods and Tools, Part 9: Quantitative Case Study Methodology," Computing and Control Engineering Journal.
8. Langford, G. O. (2012a). Engineering Systems Integration: Theory, Metrics, and Methods. Boca Raton, Florida, CRC Press, Francis and Taylor.
9. Langford, G. O. (2012b). A Theory of Systems Engineering Integration. Defence and Systems Institute (DASI), School of Electrical and Information Engineering, Mawson Lakes, Australia, University of South Australia. PhD: 331.
10. Themistocleous, M. and Irani, Z. (2002). "Towards a Novel Framework for the Assessment of Enterprise Application Integration Packages," Proceedings of the 36th Hawaii International Conference on Systems Sciences (HICSS'03), IEEE Computer Society.



11. Yin, R.K. (1981). "The Case Study Crisis: Some Answers," Administrative Science Quarterly 26, pp. 58–65.
12. Yin, R.K. (1984). Case Study Research, Saga Publications, Beverly Hills, California.
13. Yin, R.K. (1989). Case Study Research, Design and Methods. Newbury Park, Sage Publications.
14. Yin, R.K. (2003). Case Study Research. Design and Methods, 3rd Edition. Vol. 5 of Applied Social Research Method Series. Sage Publication, California.
15. Zhongyuan,Y., Rouse, W.B., and Serban, N. (2011). "A Computational Theory of Enterprise Transformation," Systems Engineering 14(4), pp.441–454.

## APPENDIX E: TRANGULATION

Triangulation—the comparison of different kinds of data (quantitative and qualitative) and different methods (e.g., observation and interviews) to see whether they corroborate one another (Silverman, 2005, p. 380).

Triangulation is one of the sampling strategy alternatives to validate the gathered data (Denzin and Lincoln, 1994). Triangulation is seen to be the answer to the dilemma of whether a specific source in the data will be robust enough to provide the conclusions of the thesis. Hence, the *combination of multiple methods, empirical materials, perspective and observers in a single study is best understood as a strategy that adds rigor, breadth, and depth to any investigation* (Denzin and Lincoln, 1994; Flick, 1992).

Triangulation can be understood from four definitions (Denzin, 1978): Data, Investigator, Theory and Methodological Triangulation; a fifth was offered by Janesick (1994): Interdisciplinary Triangulation. These triangulation factors are indicated as:

1. Data: Use of a variety of data sources in a study (Denzin and Lincoln, 1994; Janesick, 1994).
2. Investigator: Use of several different researchers or evaluators (Denzin and Lincoln, 1994; Janesick, 1994).
3. Theory: Use of multiple perspectives to interpret a single set of data (Denzin and Lincoln, 1994; Janesick, 1994).
4. Methodological: Use of multiple methods to study a single problem (Denzin and Lincoln, 1994; Janesick, 1994).
5. Interdisciplinary: Consideration of other disciplines to study a single problem (Denzin and Lincoln, 1994; Janesick, 1994).

## References:

Denzin, Norman K. 1978. The research act: A theoretical introduction to sociological methods. 2nd Edition ed. McGraw-Hill.

Denzin, NK & Lincoln, YS. (1994). "Introduction: Entering the field of qualitative research." In NK Denzin and YS Lincoln (Eds.) Handbook of Qualitative Research (pp. 1–17). Thousand Oaks: Sage Publications.

Flick, U. (1992). "Triangulation Revisited: Strategy of Validation or Alternative?" *Journal for the Theory of Social Behavior*, vol. 22, issue 2, pp. 175 – 197, June.

Janesick, V.J. (1994) "The dance of qualitative research design: Metaphor, methodolatry and meaning." Chapter 12 in N.K. Denzin & Y.S. Lincoln (Eds) *Handbook of Qualitative Research*. Sage.

## LIST OF REFERENCES

- [1] D. Nations, About.com: Web Trends, “10 great uses for Twitter: Why Twitter?” April 2012, [http://webtrends.about.com/od/twitter/a/why\\_twitter\\_uses\\_for\\_twitter.htm](http://webtrends.about.com/od/twitter/a/why_twitter_uses_for_twitter.htm).
- [2] Sprout Social, “About us,” May 2012, <http://sproutsocial.com/about>.
- [3] Work Simple, “About us,” May 2012, <http://getworksimple.com/about-us>.
- [4] Twitter Help Center, “The Twitter Glossary,” January 2012, <https://support.twitter.com/articles/166337-the-twitter-glossary>.
- [5] State of California Department of Justice, “Services and Information: California missing persons,” 2012, <http://oag.ca.gov/missing>.
- [6] N. Bruno, *Automated Physical Database Design and Tuning*. Boca Raton, FL: CRC Press, 2011.
- [7] G. Gan, *Data Clustering in C++: An Object-Oriented Approach*. Boca Raton, FL: CRC Press, 2011.
- [8] W. Trochim, *The Research Methods Knowledge Base*, 2nd Edition. Cincinnati: Atomic Dog Publishing, 2000.
- [9] Twitter Help Center, “What are hashtags (“#” symbols)?” May 2012, <https://support.twitter.com/articles/49309-what-are-hashtags-symbols>.
- [10] Commander, Navy Warfare Development Command, “Humanitarian assistance/disaster relief (HA/DR) operations planning,” Navy Warfare Development Command (NWDC) TACMEMO 3–07.6–05, August 2005.
- [11] W. Koszarek, III, “Peer production in the U.S. Navy: Enlisting Coase’s penguin,” M.S. thesis, Naval Postgraduate School, Monterey, CA, 2009.
- [12] M. Miyabe, A. Miura, and E. Aramaki, “Use trend analysis of Twitter after the great east Japan earthquake,” in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, 2012, pp. 175–178.
- [13] L. Li, S. Yang, A. Kavanaugh, E. Fox, S. Sheetz, D. Shoemaker, T. Whalen, and V. Srinivasan, “Twitter use during an emergency event: The case of the UT Austin shooting,” in *The Proceedings of the 12th Annual International Conference on Digital Government Research*, 2011, pp. 335–336.

- [14] SCLogic, "Enterprise mail & package tracking software system: Just scan, sort & deliver," May 2012, <http://www.scllogic.com/index.php>.
- [15] Klout, The Standard for Influence, "Klout measures influence online," May 2012, <http://klout.com/corp/about>.
- [16] A. Radcliffe-Brown, "On the concept of function in social science," in *Structure and Function in Primitive Society*, 1952, The Free Press of Glencoe, pp. 192–193.
- [17] D. Schoen and P. Sprague, "What Is the case method?" in *The Case Method at the Harvard Business School*, ed. M. P. McNair, New York: McGraw-Hill, 1954, pp. 78–79.
- [18] R. Yin, *Applications of case study research*. Beverly Hills: Sage Publishing, 1993.
- [19] J. Meyers, "How to track down a celebrity," eHow.com, [http://www.ehow.com/how\\_4457489\\_track-down-celebrity.html](http://www.ehow.com/how_4457489_track-down-celebrity.html).
- [20] Newport News, Va., Daily Press, "Poquoson girl's prom date campaign leads to awards show invite from pop star Justin Bieber," 20 May 2012, <http://www.dailypress.com/entertainment/dp-nws-bieber-poquoson-20120520-14,0,3023878.story>.
- [21] L. Altman, South Bay Crimes and Courts, "LAPD announces milestone number of officers," March 2009, <http://www.insidesocal.com/crime&courts/2009/03/lapd-announces-milestone-numbe.html>.
- [22] CNN Wire Staff, CNN U.S., "Dodgers challenge lawsuit filed by family of beaten fan," August 2011, [http://articles.cnn.com/2011-08-12/us/california.baseball.beating\\_1\\_bryan-stow-security-officers-frank-mccourt?s=PM:U.S.](http://articles.cnn.com/2011-08-12/us/california.baseball.beating_1_bryan-stow-security-officers-frank-mccourt?s=PM:U.S.).
- [23] R. Merton, *Social Theory and Social Structure*. New York: The Free Press, Simon & Schuster, 1968.
- [24] Z. Chu, S. Gianvecchio and H. Wang, "Who is tweeting on Twitter: Human, Bot, or Cyborg?" in *Proceedings of the 26th Annual Computer Security Applications Conference*, 2010, pp. 21–30.
- [25] Twitter Help Center, "Guidelines for law enforcement" May 2012, <http://support.twitter.com/articles/41949-guidelines-for-law-enforcement>.
- [26] Infographic Labs, "Twitter 2012," February 2012, <http://infographiclabs.com/news/twitter-2012/>.

- [27] A. Ostrow, Mashable Social Media, "Social networking dominates our time spent online [STATS]," 2 August 2010,  
<http://mashable.com/2010/08/02/stats-time-spent-online/>, 6 March 2012
- [28] Google Analytics Blog, "Capturing the value of social media using google analytics," 20 March 2012,  
<http://analytics.blogspot.com/2012/03/capturing-value-of-social-media-using.html>.
- [29] D. Williamson, Ad Age Digital, "How much will you spend on social-media marketing next year?," 8 December 2010,  
<http://adage.com/article/digitalnext/social-media-marketing-spend-year/147544/>.
- [30] Twitter, "Terms of service," May 2012, <https://twitter.com/tos>.
- [31] Twitter, "Twitter privacy policy," May 2012, <http://twitter.com/privacy>.
- [32] Twitter Blog, "Twitter turns six," 21 March 2012,  
<http://blog.twitter.com/2012/03/twitter-turns-six.html>.
- [33] Twitter Developers, "Getting started," May 2012,  
<https://dev.twitter.com/start>.
- [34] M. Pradel and T. Gross, "Mining API usage protocols from large method traces," in *Mining Software Specifications: Methodologies and Applications*, L. Chao, Ed. Boca Raton: CRC Press 2011, Ch. 4.
- [35] N. Kho, "Location, location, (geospatial) location," *Information Today*, vol. 27, p. 1, July/August 2010.
- [36] Twitter Blog, "Location, location, location," August 2009,  
<http://blog.twitter.com/2009/08/location-location-location.html>.
- [37] J. Cowie and Y. Wilks, "Information extraction," in *Handbook of Natural Language Processing*, R. Dale, H. Moisl, and H Somers, Ed. New York: Marcel Dekker, Inc., 2000, Ch. 10.
- [38] C. Samuelsson, M. Wir`en, "Parsing techniques," in *Handbook of Natural Language Processing*, R. Dale, H. Moisl, and H Somers, Ed. New York: Marcel Dekker, Inc., 2000, Ch. 4.
- [39] B. McLaughlin, *Java and XML*. Sebastopol: O'Reilly Media, 2000.
- [40] D. Ferrara, About.com: Web Design/HTML, "All about XML parsers – what do XML parsers do: XML parser F.A.Q.," April 2012,  
<http://webdesign.about.com/od/parsers/a/all-about-xml-parsers.htm>.

- [41] B. Thuraisingham, *XML Databases and the Semantic Web*. Boca Raton: CRC Press LLC, 2002.
- [42] J. Atwood, Coding Horror, "Parsing html the Cthulhu way," November 2009, <http://www.codinghorror.com/blog/2009/11/parsing-html-the-cthulhu-way.html>.
- [43] K. Gladdis, Mail Online: Science Home "Twitter secrets for sale: Privacy row as every tweet for last two years is bought up by data firm," February 2012, <http://www.dailymail.co.uk/sciencetech/article-2107693/Twitter-sells-years-everyones-old-vanished-tweets-online-marketing-companies.html>.
- [44] RT Question More, "Privacy betrayed: Twitter sells multi-billion tweet archive," February 2012.
- [45] L. Chao, *Database Development and Management*, Auerbach Publications, 2006.
- [46] C. Fehily, *SQL: Visual QuickStart Guide*, 3<sup>rd</sup> ed., Berkeley: Peachpit Press, 2008.
- [47] A. Oppel and R. Sheldon, *SQL: A Beginner's Guide*, 3<sup>rd</sup> ed. McGraw-Hill, 2008.
- [48] P. Murrell, *Introduction to Data Technologies*. Boca Raton: Chapman and Hall, 2009.
- [49] w3schools, "SQL wildcards," May 2012, [http://w3schools.com/sql/sql\\_wildcards.asp](http://w3schools.com/sql/sql_wildcards.asp).
- [50] A. Joshi and R. Motwani, "Keyword generation for search engine advertising," in *International Conference on Data Mining Workshops*, 2006, pp. 490–496.
- [51] C. Sherman and G. Price, *The Invisible Web: Uncovering Information Sources Search Engines Can't See*. Medford: CyberAge Books, 2001.
- [52] A. Fuxman, P. Tsaparas, K. Achan, and R. Agrawal, "Using the wisdom of the crowds for keyword generation," in *Proceedings of the 17th International Conference on World Wide Web*, 2008, pp. 61–70.
- [53] R. Ho, *Handbook of Univariate and Multivariate Data Analysis and Interpretation with SPSS*. Boca Raton: Chapman & Hall/CRC, 2006.
- [54] J. Huang, K. Thornton, and E. Efthimiadis, "Conversational tagging in Twitter," in *Proceedings of the 21<sup>st</sup> ACM Conference on Hypertext and Hypermedia*, 2010, pp. 173–178.

- [55] Press Release, Directions Magazine, "Geosemble release integrates social media with the latest satellite imagery," 10 May 2012, <http://www.directionsmag.com/pressreleases/geosemble-release-integrates-social-media-with-the-latest-satellite-im/251699>.
- [56] CBC News, "Halifax regional police use Twitter to fight crime," 28 February 2012, <http://www.cbc.ca/news/canada/nova-scotia/story/2012/02/28/ns-halifax-regional-police-twitter.html>.
- [57] *Los Angeles Times*, "L.A. arson: A turning point for police use of Twitter, social media," L.A. Now section, 2 January 2012, <http://latimesblogs.latimes.com/lanow/2012/01/la-arson-fires-police-use-twitter-social-media-twitter-social-media.html>.
- [58] T. Heverin and L. Zach, "Twitter for city police department information sharing," in *Proceedings of the 73<sup>rd</sup> ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem*, 2010, Volume 47, Article No. 41.



THIS PAGE INTENTIONALLY LEFT BLANK

## INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center  
Ft. Belvoir, Virginia
2. Dudley Knox Library  
Naval Postgraduate School  
Monterey, California
3. Gary Langford  
Naval Postgraduate School  
Monterey, California